3-29-2021 3:00 PM

A Deep Topical N-gram Model and Topic Discovery on COVID-19 News and Research Manuscripts

Yuan Du, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science

Abstract

Topic modeling with the *latent semantic analysis* (LSA), the *latent Dirichlet allocation* (LDA) and the *biterm topic model* (BTM) has been successfully implemented and used in many areas, including movie reviews, recommender systems, and text summarization, etc. However, these models may become computationally intensive if tested on a humongous corpus. Considering the wide acceptance of machine learning based on deep neural networks, this research proposes two deep neural network (NN) variants, 2-layer NN and 3-layer NN of the LDA modeling techniques. The primary goal is to deal with problems with a large corpus using manageable computational resources.

This thesis analyze two datasets related to COVID-19 to explore the underlying structures. The first dataset includes over 7,000 CBC COVID-19 related news articles for the period of January 9, 2020 to May 3,2020. The second dataset, called CORD-19, includes over 100,000 research manuscripts related to COVID-19 for the period of January 2, 2020 to August 1, 2020. We discovered that in the first dataset 14 topics were including “traveling”, “lockdown”, “masks”, the focus of social media attention during the period of January to May of 2020. For the second dataset, 17 topics, including "vaccine", “treatment” and "social distancing", were identified to be the focus of research articles for the period of January to August of 2020. Compared to the traditional LDA, our proposed model requires less computation time and shows better performance.

Keywords

COVID-19, N-gram, LDA, Topic Modeling, Machine Learning, Deep Learning

Summary of Lay Audience

Topic modeling is an unsupervised machine learning technology that detects the structures of words and phrases in documents. It is one of the most powerful techniques for text mining. Topic modeling provides us with methods to organize, understand and summarize large amounts of textual information. It helps us to discover hidden theme patterns that exist in the collection and annotate documents. Topic modeling has been widely used in applications, and a large number of articles have been published in various fields such as software engineering, political science, medical science, and linguistics. For example, topic modeling has been applied to analyze information collected by social media websites such as Twitter and Facebook.

In this thesis, we analyze two datasets related to COVID-19. We use the Natural Language Processing (NLP) methods to identify topics and keywords related to COVID-19 from the news and research manuscripts. We discovered that in the first dataset 14 topics were including “traveling”, “lockdown”, “masks”, the focus of social media attention during the period of January to May of 2020. For the second dataset, 17 topics, including "vaccine", “treatment” and "social distancing", were identified to be the focus of research articles for the period of January to August of 2020. Compared to the traditional LDA, our proposed model requires less computation time and shows better performance.

List of Key Acronyms Throughout the Thesis

Acronyms	Description
AI	Artificial Intelligent
BOW	Bag of Words
BTM	Bi-term Topic Model
CBC	Canadian Broadcasting Corporation
CORD	COVID-19 Open Research Dataset
DTNG	Deep Topical N-gram Model
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
NN	Neural Network
SSE	Sum of Squared Errors
TF-IDF	Term Frequency and Inverse Document Frequency
TNG	Topical N-gram Model

Acknowledgments

I wish to express my deepest gratitude to my supervisor and thesis committee members, Professor Boyu Wang, Professor Michael Bauer and Professor XianBin Wang, without whom this work would have not come together. Their kind feedback and comments were the best guide throughout the journey, and I learnt many great things from them not only as an academic supervisor but as a person.

I would like to pay my special regards to my parents and friends, who like always to have been supportive.

I would like to offer my gratitude and thanks to everyone who supported me in whatever way in my graduate studies at UWO.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Topic modeling is an unsupervised machine learning technology that can detect words and phrases in a set of documents and is one of the most powerful techniques for text mining, latent data discovery, and finding relationships among data and text documents [1]. It is essentially defined as a statistical method of text mining, used to identify potential (hidden) patterns in a data corpus and classify key words in a corpus as topics.

The use of topic modeling is very wide. Researchers have published many articles in various fields such as software engineering, political science, medical science, and linguistics. For example, topic modeling based on social media analysis helps us understand reactions and conversations among people on the Internet; it extracts useful information from the reactions and content shared on social media websites such as Twitter and Facebook [2-3]. Topic modeling can also be defined as a scientific method for tracking word clusters, called topics, in large amounts of text. In statistical terms, a topic can be loosely treated as a multinomial distribution of words that appear simultaneously. By running the topic model iteratively on a given collection, topics can be learned from the document collection.

Using data based on CBC News and COVID-19 research papers, in this thesis, we analyze news and text related to COVID-19. Specifically, using cluster-based natural language processing (NLP) methods, we extract information from the datasets to discover various topics and keywords related to COVID-19. Moreover, we propose to apply the neural network to the existing LDA method to improve the performance of model fitting.

This chapter is organized as follows. In Section 1.1, we briefly discuss COVID-19 disease and its related research articles. Our motivation, contribution for this thesis and the structure of this thesis are described in Section 1.2.

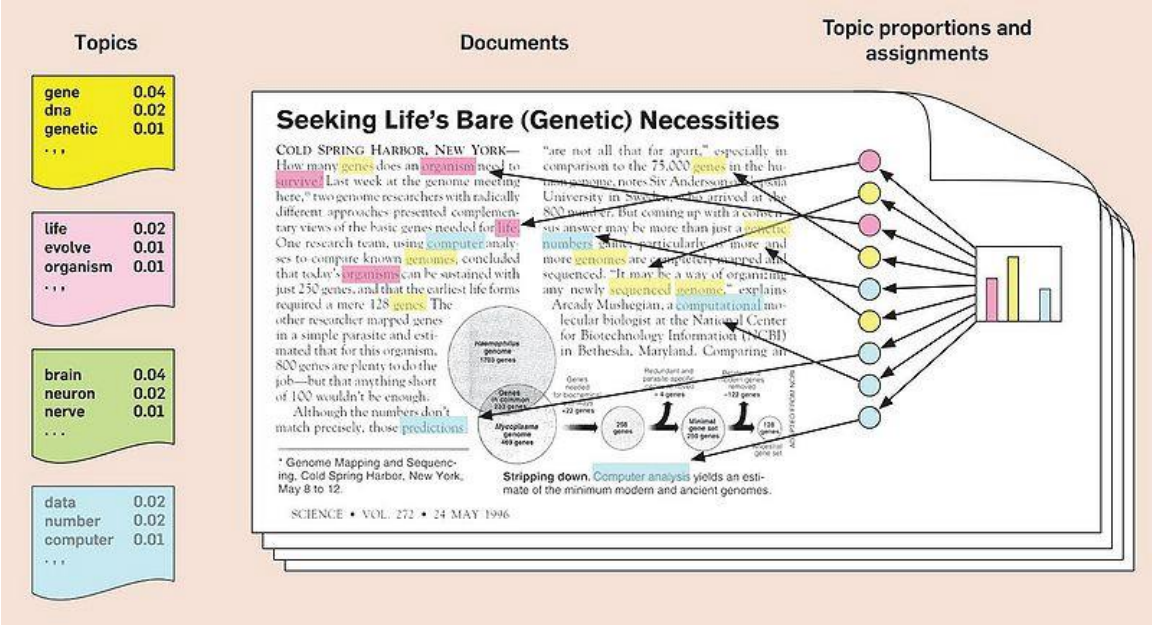

Figure 1: Illustration of Probabilistic Topic Modeling [62]

1.1 Coronavirus disease (COVID-19) and Related Articles

Coronaviruses are a large family of viruses that cause diseases such as the Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV). The novel coronavirus disease (COVID-19) is a new species that was discovered in 2019 and has not been previously identified in humans. Coronaviruses are zoonotic due to contamination from animals to humans [4].

The outbreak of COVID-19 was reported in late December 2019 [1]. COVID-19 was first discovered in Wuhan, China, and has spread rapidly worldwide, resulting in a global pandemic since March 2020 [2]. Due to the rapid spread of the virus, the World Health Organization declared a state of emergency on March 11. More than 11 million cases and 545,000 deaths have been recorded in more than 200 countries and regions as of July 1, 2020 [3].

Here, we are interested in investigating what the Canadian news was reporting on COVID-19 during the pandemic, and what researchers were paying attention to about COVID-19 through writing articles. To do that, we extracted the number of confirmed, death, and recovered cases in Canada from the official website of the Government of

Canada [5]. Figure 2 shows the number of confirmed, death, and recovered cases in Canada for the period of February 1 to July 1 , 2020.

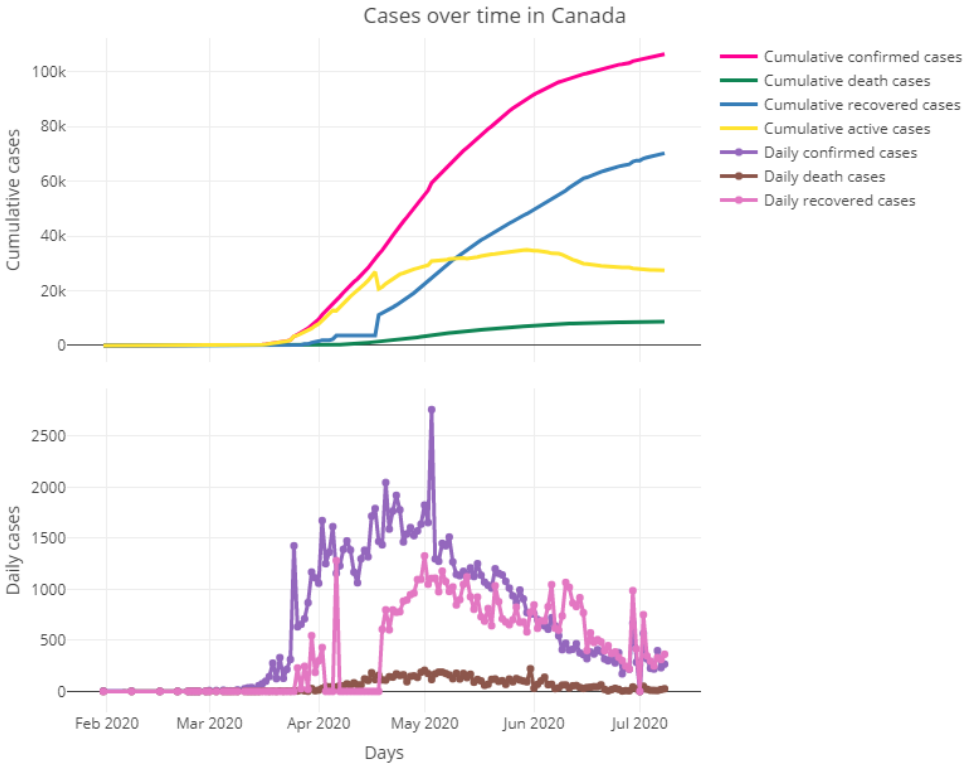

Figure 2: The number of COVID-19 confirmed, death, and recovered cases in Canada [3]

Many researchers have used deep learning methods to analyze COVID-19 data with social media such as Twitter and reddit [6] - [11]. Just to name a few, Kouzy and colleagues [12] focused on analyzing the severity of misinformation about the spread of the COVID-19 epidemic on social media. They conducted a search on Twitter using 14 different trending hashtags and keywords related to the COVID-19 epidemic. Medical misinformation and unverifiable content about the global COVID-19 epidemic spread on social media at an alarming rate. The importance of intervening in the dissemination of misinformation was emphasized in order to curb this phenomenon that endangers public safety when awareness and appropriate preventive actions are essential. Kreps and Kriner [13] conducted two studies to examine the pervasiveness and persuasiveness of misinformation about the origin of the new coronavirus, effective treatments, and the effectiveness of government responses. In all categories, they found that even well-

known false claims have relatively low true recall rates. They found little evidence that exposure to misinformation would seriously affect a series of policy beliefs and political judgments. Nanning and colleagues [14] proposed a hybrid artificial intelligence (AI) model for COVID-19 prediction, which embeds natural language processing (NLP) modules and long short-term memory (LSTM) networks into the ISI model to build a hybrid for COVID-19 prediction. Their study shows that the openness, transparency, and efficiency of releasing data are very important for establishing a modern epidemic prevention system. Hamed and colleagues [15] used a *Long Short Terms Memory* (LSTM) Recurrent Neural Network Approach to detect meaningful latent-topics and sentiment-comment-classification on COVID-19–related issues from healthcare forums and social media. Pedram and colleagues [16] collected more than 530,000 original tweets in Persian/Farsi related to COVID-19 pandemic over time and analyzed the content in terms of major topics of discussions and a broader category of tweets.

1.2 Motivations and Contributions

Topic modeling is a useful and effective technology in *Natural Language Processing* (NLP). It is mainly used for semantic mining and latent discovery in documents and datasets. In this thesis, we focus on analyzing emotions and semantic thoughts related to COVID-19 based on CBC news and research manuscripts. Using the Natural Language Processing (NLP) method based on clustering from the dataset, we extract various topics and keywords related to COVID-19 from the news and manuscripts to analyze the focus of news and research from January to May, 2020.

A good topic model is believed to identify a meaningful cluster of words. Cluster of words that can be unambiguously labeled as a meaningful representation of some common topic. Common topic modeling techniques include PLSI [26], LDA and Biterm Topic Model (BTM) [27]. Here we consider using deep neural networks to combine with the existing topical N-gram model. The proposed models are believed to reduce the computational cost required in LDA to extract topics (themes) of a huge corpus.

This thesis is organized as follows. First, we provide a brief introduction to text classification and clustering. Discussion of COVID-19 related issues and some relevant

work are provided in Chapter 2. In Chapter 3, we describe the COVID-19 data and pre-processing methods adopted in our research. Chapters 4 and 5 present our experiment settings for the two COVID-19 related datasets. The results and discussions are reported in Chapter 6. Finally, we summarize our work and discuss future work in Chapter 7.

Chapter 2 Related Work and Literature Review

In this chapter, we introduce serval methods related to topic modeling. In Section 2.1, we present text classification and its related works. We introduce clustering techniques in Section 2.2. Section 2.3 shows the TF-IDF method which is used to evaluate the importance of words in the document. In Section 2.4, we describe the N-gram model, a useful model in text classification. We explain the LDA method in detail in Section 2.5. Sections 2.6 - 2.8 present two N-gram models, the bi-gram topic model and the topical N-gram model. Finally, in Section 2.9, we describe our proposed model.

2.1 Text Classification

Text classification automatically classifies text sets (or other entities or objects) according to a certain classification system or principle. It finds the relationship model between document features and document categories based on a set of labeled training documents, and then uses this learned relationship model to classify new documents [17]. Figure 3 shows the graphical representation of the text classification process. A classifier takes the text as an input, then changes each word to lowercase and tokenizes it to single word, deletes the stop words (i.e., “in”, “where”, “to”, etc.), and uses the feature extractor to get the vectors. These new vector sets can aggregate most of the information contained in the original feature set. Then one applies those vectors to a machine learning algorithm, followed by the application of a classifier model.

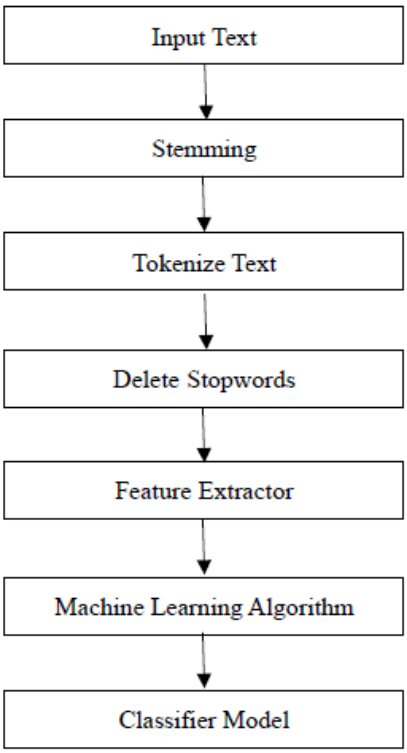

Figure 3: The Text Classification Process

A common method of text classification is to create an instance-level representation by using the final hidden state of the RNN, the maximum (or average) pool of hidden states of the RNN, or convolutional N-grams. However, such a method may ignore the importance of special words that are highly distinguishable for classification. Bahdanau et al. [50] introduced the attention mechanism in machine translation, which has been used in many natural language processing tasks. Yang et al. [51] considered an integral part of the model for text classification. Lin et al. [52] proposed a new model for extracting interpretable sentence embeddings using self-attention. Ma et al. [53] showed that the attention mechanism is also effective for sentiment classification. Vaswani et al. [54] further illustrated that a stronger sentence-level representation can be obtained by stacking multiple blocks of self-attention. Bert et al. [55] combined the transformer and a large corpus to produce a more complete and better sentence-level representation.

The word matching or word searching method is the earliest classification algorithm. This method only judges whether the document belongs to a certain category based on

whether there is a word with the same name in the document (the processing of adding synonyms at most) [18]. Obviously, an over-simple mechanical method cannot bring in a good classification outcome. The method of knowledge engineering has emerged later on. With the help of professionals' experts, one may define a large number of inference rules for four different types of knowledge: (i) domain knowledge, (ii) inference knowledge, (iii) task knowledge, and (iv) strategic knowledge [19]. If a document can meet these inference rules, it can be determined to belong to a certain category. With the inclusion of human judgment to the system, the accuracy is greatly improved. However, the shortcomings of this method are still obvious. The most fatal weakness of knowledge engineering lies in its lack of generalizability. A classification system built for the financial field cannot be extended to related fields such as medical or social insurance. Afterwards, text classification has gradually changed from a knowledge-based method to a method based on statistics and machine learning.

2.2 Clustering

Clustering techniques are part of text classification [20]. Clustering is the task of dividing the population or data points into a number of groups so that data points in the same groups are more similar to those data points in the same group and dissimilar to the data points in other groups.

Cluster analysis itself is not a specific algorithm, but a general task to be solved. It can be implemented by various algorithms that differ greatly in understanding what constitutes clusters and how to find them efficiently [21]. There are four major steps in clustering [59]:

- Feature selection or extraction: The process of determining which attributes of data objects are used to distinguish objects. In this process, extraction can derive new features from existing feature attributes.

- Algorithm design: Determine proximity and construct a criterion function. If the data objects are similar or different according to their functions, they are intuitively divided into different groups.

- Cluster verification: There are usually three types of verification standards: external testing, internal indicators, and relative indicators. These three types are defined as three types of cluster: partition clusters, hierarchical clusters and 3 separate clusters.

- Interpretation of results: Give meaningful data insights through results.

Popular concepts of clustering include small distances between cluster members, dense regions of data space, and intervals or groups of a specific statistical distribution. Clustering can be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including parameters, such as the distance function to be used, the density threshold, or the number of expected clusters) depend on the individual data set and the intended use of the results. Such a cluster analysis is not an automatic task, but an iterative process involving knowledge discovery or interactive multi-objective optimization of discovery and trial. It is usually necessary to modify the data preprocessing and model parameters until the obtaining the required attributes.

2.3 Term Frequency and Inverse Document Frequency

Term Frequency and Inverse Document Frequency (TF-IDF) is used to evaluate the importance of words to a document in the document collection [22]. TF is the short for “term frequency” and IDF is the short for “inverse document frequency”. The TF-IDF algorithm assumes that the most meaningful words for differentiating documents should be those that appear frequently in the document and less frequently in other documents in the entire document collection. Therefore, if the feature space uses the TF word frequency as a metric, it reflects the features of similar texts. In addition, considering the ability of words to distinguish between different categories, the TF-IDF method assumes that the smaller the frequency of text in which a word appears, the greater its ability to distinguish between different categories of text. Therefore, the concept of inverse document frequency IDF is introduced. The product of TF and IDF is used as the measurement of the coordinate system of the feature space, and it is used to complete the adjustment of the weight TF. The purpose of adjusting the weight is to highlight important words and suppress secondary words. But in essence, IDF is a weighted attempt to suppress noise, and simply thinks that words with a small text frequency are

more important, and words with a large text frequency are less useful or even useless. Obviously, this is not completely correct. The simple structure of IDF cannot effectively reflect the importance of words and the distribution of characteristic words, making it unable to complete the function of adjusting weights well [23].

By multiplying the term frequency of a word in a document and the inverse document frequency of the word across a set of documents results in the TF-IDF score of a word in a document. The higher the score, the more relevant that word is in that particular document. To put it in more formal mathematical terms, the TF-IDF score for the word t in the document d from the document set D is calculated as follows:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D),$$

where

$$tf(t,d) = log\{1 + freq(t,d)\};$$

$$idf(t,D) = log\left\{\frac{N}{count(d \in D: t \in d)}\right\}.$$

Here, $freq(t,d)$ represents the document frequency of the term t, $count\ (d \in D: t \in d)$ is the number of times that term t appears in document d, and N is the total number of documents.

The TF-IDF algorithm is simple and fast to implement, and the result is likely in line with the actual situation. Its drawback is that simply measuring the importance of a word with "word frequency" is not adequate enough. Sometimes important words may not appear frequently. Moreover, this algorithm cannot reflect the position information of the words, and the words appearing in different positions in the document are regarded to have the same importance, which may not be true. On the other hand, the dimension of the vectors can be high due to the large list of vocabularies, thus needing a very large memory to encode the document, which slows down the speed of the algorithm. Because of these drawbacks, TF-IDF is often just used as a baseline model to evaluate newly developed word embedding model.

2.4 N-gram Language Model

The *N-gram Language Model* is a language model developed by Stanford University in 2019 [28]. N-gram modeling is a feature recognition and analysis method used in language modeling and natural language processing. N-gram is a sequence of consecutive items of length N, which can be a sequence of words, bytes, syllables, or characters. The most commonly used N-gram models in text classification are word-based and character-based N-grams. In this thesis, we use word-based N-grams to represent the context of documents and generate functions to classify documents. The N-gram model can be presented as:

$$P(w|y) = \frac{P(w)P(y|w)}{P(y)}.$$

In this model, w is a word-series hypothesis representing a series of one or more words, for example English-language words. The term $P(w)$ is the probability of occurrence of the word-series hypothesis w. The variable y is an observed signal, and $P(y)$ is the probability of occurrence of the observed signal y. $P(w|y)$ is the conditional probability of occurrence of the word series w, given the occurrence of the observed signal y. $P(y|w)$ is the conditional probability of occurrence of the observed signal y, given the occurrence of the word-series w.

2.5 Latent Dirichlet Allocation (LDA)

LDA, standing for *Latent Dirichlet Allocation*, was proposed by David Blei, Andrew Ng and Michael O. Jordan in 2003 [41]. It is a generative probability model used to collect discrete data, such as text corpus, genome sequences, and image collections, etc. LDA can be used to model discrete data in any domain, but in order to explain its generation process, we here pay special attention to text collection. In text collections, LDA explores the possibility of representing documents as a random mixture of potential topics, where each topic is a multinomial distribution of words. When implementing LDA, we need to have a collection D of documents d, where each $d = \{w_1, \dots, w_N\}$ with w_i representing the i-th word in the sequence for $i = 1, \dots N$.

LDA has a purely probabilistic view of the document generation process. The two polynomial distributions, usually expressed as ϕ_t and ϕ_d, are pivotal in LDA. The basic assumption about LDA is that each word ω_i extracted from the document collection may be associated with a potential topic, say z_k. The probability can be approximated by using the polynomial distribution ϕ_t on the vocabulary V of the corpus, such that

$$P(\omega_i|z_k = t) = \phi_t(z_k, w_i).$$

The assumption is that each document mixes with various topics and every topic mixes with various words. Table 1 shows an example of various words under various topics. Intuitively, we can image that we have two layers of aggregations. The first layer is the distribution of categories. For example, we have finance news, weather news and political news. The second layer is the distribution of words within the category. For instance, we can find "sunny" and "cloud" in weather news while "money" and "stock" in finance news.

Table 1: Examples of various words under various topics

Finance	**Weather**	**Arts**	**Sport**
Money	Sunny	Music	World Cup
Stock	Cloud	Piano	Soccer
Trade	Humanity	Calligraphy	Tennis
Market	Temperature	Photography	Sailing

However, "a", "with" and "can" do not contribute useful information on topic modeling. Those words exist among documents and usually have roughly the same probability across categories. Therefore, stop words removal is a critical step to achieve a better result.

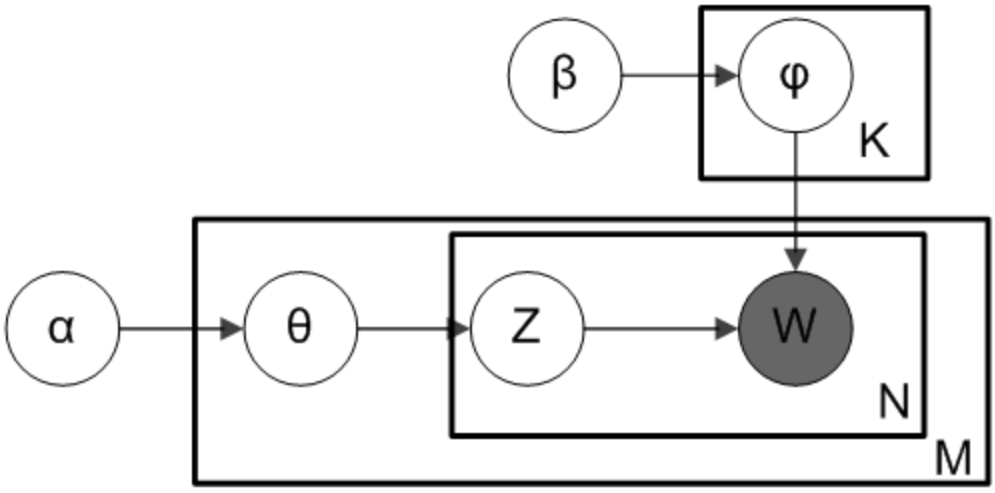

Figure 4: LDA Model [41]

Figure 4 represents an LDA model. The boxes are "plates" representing replicates, which are repeated entities. The outer plate represents documents, while the inner plate represents the repeated word positions in a given document; each position is associated with a choice of topic and word [46]. The variable names in Figure 4 are defined as follows:

- M denotes the number of documents
- N is the number of words in a given document (e.g., document d has N_d words)
- K is the number of topics
- α is the parameter of the Dirichlet prior on the per-document topic distributions
- β is the parameter of the Dirichlet prior on the per-topic word distribution
- θ_d is the topic distribution for document d
- φ_t is the word distribution for topic t
- z_{dn} the topic for the n-th word in document d
- w is the observed word.

The letter W in Figure 4 is grayed to indicate that words w_{ij} are the only observable variables, and other variables are latent.

Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution of all the words. Formally, the process for generating each word from a document is as follows:

1. Select Discrete distributions θ_d for document d ($d \in \{1 \dots, M\}$) from a Dirichlet distribution with parameter α.
2. Select Discrete distributions ϕ_k for topic t ($t \in \{1 \dots, K\}$) from a Dirichlet distribution with parameter β.
3. For a word w_n ($n \in \{1, \dots, N_d\}$) in document d,
    1) Select a topic z_n from a Discrete distribution θ_d.
    2) Select a word w_n from a Discrete distribution φ_{zn}.

In the above generative process, words (w) in documents are the only observed variables, while others are latent variables (φ and θ) or hyper parameters (α and β). According to the model, the probability of observed data D is computed by

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\sum_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w|z_{dn}, \varphi_t) \mathrm{P}(\varphi_t|\beta) \right) d\theta_d d\varphi_t \, .$$

2.6 Bigram Topic Model

The Bigram Topic Model (BTM) employs the Hierarchical Dirichlet language model [27] by incorporating the concept of the topic into the bigram model.

For this model, we assume a dummy word w_0 existing at the beginning of each document. The graphical model presentation of this model is shown in Figure 5 where M documents are considered. The generative process of this model is described as follows:

For a given document,

1. Select Discrete distributions σ_{zw} from a Dirichlet prior δ for each topic z and each word w.

2. Select a Discrete distribution θ from a Dirichlet prior α.

3. For each word w_n in document:

    a) select z_i from a Discrete distribution θ; and

    b) select w_i from Discrete distributions $\sigma_{z_i w_{i-1}}$

Here, $\sigma_{z_i w_{i-1}}$ means the multinomial (Discrete) bigram distribution of words with respect to topic z with the previous word w.

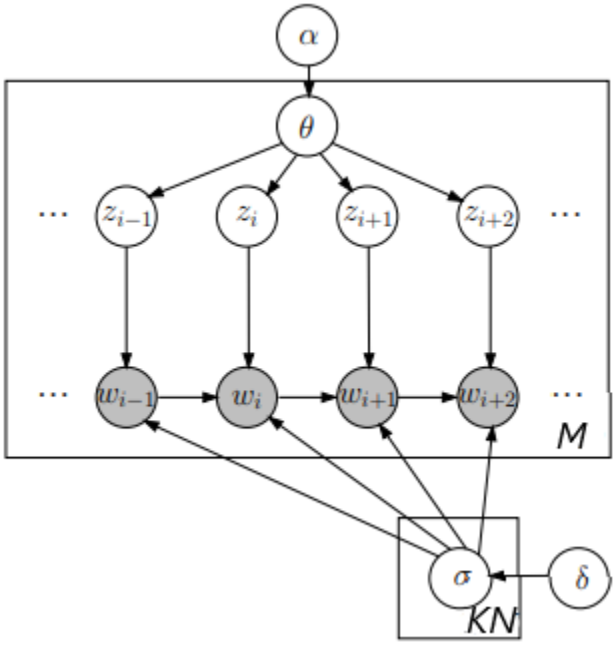

Figure 5: Bigram Topic Model [27]

2.7 Topical N-gram Model

The *topical N-gram* (TNG) model [47] is a combination of the Bigram Topic model and the LDA Collocation model. It solves the problem associated with keywords as in the Bigram Topic model, and automatically determines whether a composition of two terms is indeed a bigram as in the LDA collocation model. However, like other collocation discovery methods, a discovered bigram is always a bigram in the LDA Collocation model. One of the main contributions of this model is to make it possible to decide whether to form a bigram for the same two consecutive word tokens depends on their nearby context (i.e., cooccurrences). As in the LDA collocation model, we may assume

some of x are observed for the same reason, where x represents random variables for bigram status. The graphical model presentation of this model is shown in Figure 6, and its generative process is described as follows:

1. Select Discrete distributions φ_z from a Dirichlet prior β for each topic z;
2. Select Bernoulli distributions ψ_{zw} from a Beta prior γ for each topic z and each word w;
3. Select Discrete distributions σ_{zw} from a Dirichlet prior δ for each topic z and each word w;
4. For each document, select a Discrete distribution θ from a Dirichlet prior α;
5. For each word w_i in the document :
    a) Select x_i from Bernoulli distributions $\psi_{z_{i-1}w_{i-1}}$;
    b) Select z_i from a Discrete distribution θ;
    c) Select w_i from Discrete distributions $\sigma_{z_i w_{i-1}}$ if $x_i = 1$; else select w_i from Discrete distributions ϕ_{z_i}.

Here, x_i represent the bigram status between the $(i-1)$-th token and i -th token in the document. γ is the Dirichlet prior of ψ, and ψ is the binomial (Bernoulli) distribution of status variables with respect to topic z and word w.

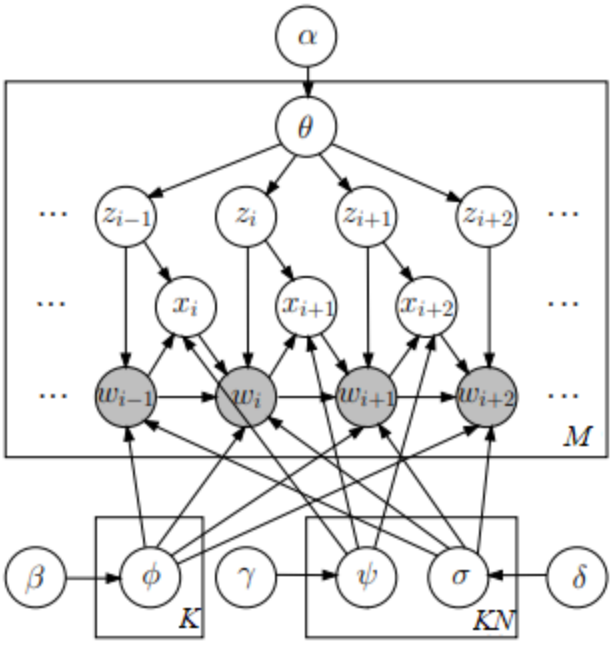

Figure 6: Topical N-gram Model [47]

2.8 Topic Modeling

In text mining, some documents need to be collected, such as blog posts or news articles, and then divided into natural groups so that people can understand them separately and clearly. Topic modeling is a graphical model that automatically analyzes text data to determine cluster words for a set of documents. This is known as ‘unsupervised’ machine learning because it does not require a pre-defined list of tags or training data that have been previously classified by humans. Since topic modeling does not require training, it is a quick and easy way to start for analyzing data.

All topic models make the same basic assumption:

- each document consists of a mixture of topics, and
- each topic consists of a collection of words.

In other words, the topic model is constructed based on the idea that the semantics of the document is actually controlled by some unobserved hidden or "latent" variables.

Therefore, the goal of topic modeling is to discover these latent variables (topics) that suggest the meaning of documents and corpus.

Topic modeling can be easily compared to clustering. As in the case of clustering, the number of topics, like the number of clusters, is a hyperparameter. By doing topic modeling we build clusters of words rather than clusters of texts. A text is thus a mixture of topics, each having a certain weight. If document classification is assigning a single category to a text, topic modeling is assigning multiple tags to a text. A human expert can label the resulting topics with human-readable labels and use different heuristics to convert the weighted topics to a set of tags. [61]

There are several algorithms associated with topic modeling. The most popular ones include Latent Semantic Analysis (LSA) [33], Latent Dirichlet Allocation (LDA), and Biterm Topic Model (BTM). Because LSA focuses on reducing matrix dimension, BTM is used for short texts, and LDA solves topic modeling problem. In this thesis, we only consider LDA.

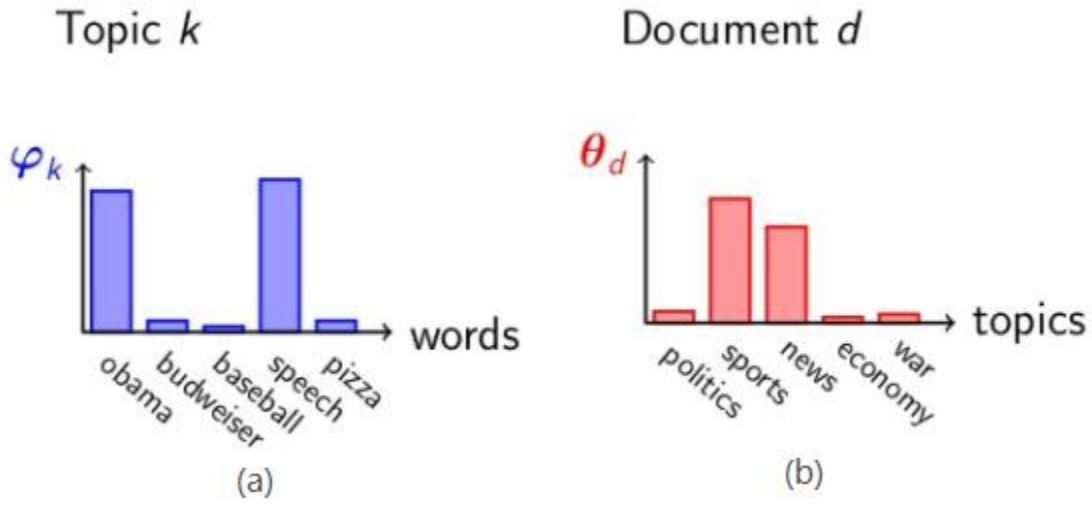

Figure 7: (a)Identification of a collection of documents using LDA.

(b) Tagging each document with topics using LDA.

The idea of using collocations in the topic model is not new, which basically is to create a unified probability model with an initial extraction of collocations and N-grams and further integrated into the topic model.

The first initiative beyond the "bag-of-words" assumption was made by Wallach [56], where the Bigram Topic Model was presented. In this model word probabilities are conditioned on the immediately preceding word. The LDA Collocation Model [57] extends the Bigram Topic Model by introducing a new set of variables and thereby giving the flexibility to generate both unigrams and bigrams. Wang et al. [47] proposed the Topical N-Gram Model that adds a layer of complexity to allow the formation of bigrams to be determined by the context. Hu et al. [58] proposed the Topical Word Character Model, relaxing the assumption that the topic of an N-gram is determined by the topics of composite words within the collocation. This model is mainly suitable for Chinese language. Johnson [59] established the connection between LDA and Probabilistic Context-Free Grammars, and proposed two probabilistic models by combining insights from LDA and Adaptor Grammars to integrate collocations and proper names into the topic model.

2.9 Deep Topical N-gram Model

In this section, we propose two sequential deep topical N-gram models, called 2NN DTNG and 3NN DTNG. We use Keras, an open-source neural network library, with two or three hidden layers to extend traditional Topic Modeling techniques. The input layer size of the proposed deep models (2NN and 3NN DTNG) is equal to the size of the vocabulary extracted from the given dataset remained after prerequisite pre-processing.

The Keras sequential model with two hidden layers and three hidden, as mentioned in Figure 8 is created with following specifications.

a) The input layer with the number of nodes equal to the size of vocabulary (length of dictionary) extracted from CORD-19.

b) Activation function, loss function, and optimizer are respectively chosen as 'tanh', 'categorical_crossentropy', and 'sgd'.

c) Both models (2NN DTNG and 3NN DTNG) are trained in 100 epochs. Models are saved for subsequent use.

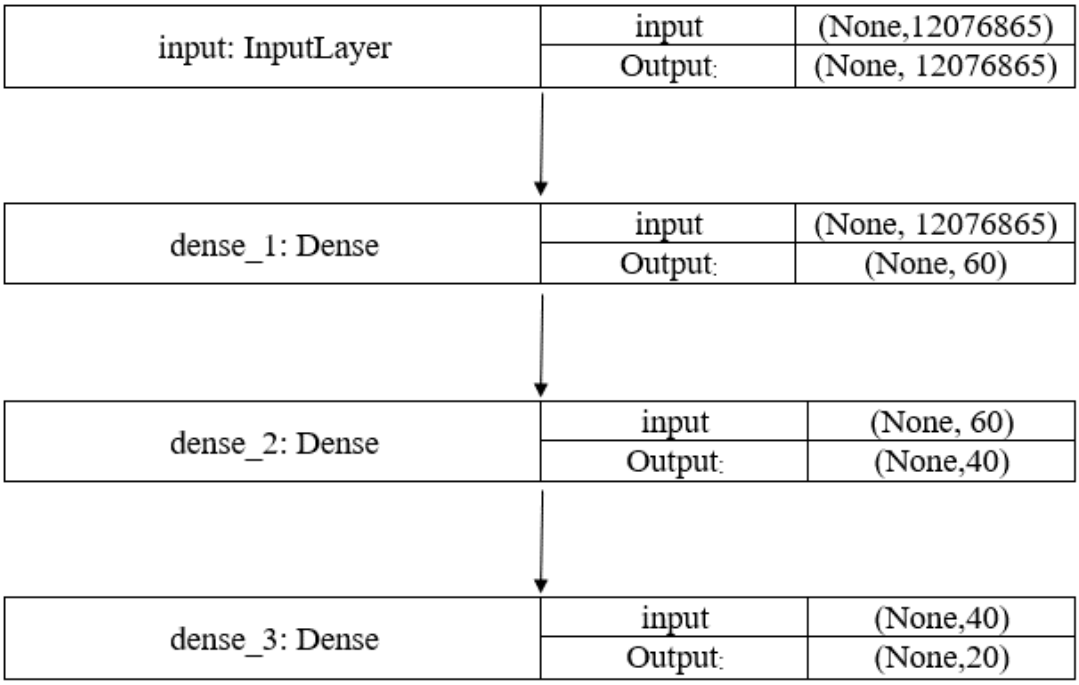

Figure 8: Model Summary of the proposed 3NN DeepTNG with CORD-19 dataset

Vocabulary extracted from the dataset is stemmed by using NLTK's WorldNet Stemmer. We take the CORD-19 dataset as an example. Input documents were first converted into *bag of word* (BOW) [48] representation, which includes 12,076,865 vectors. We first put the vectors into the TNG model for training, and then use the training results as a supervisor for neural network training. As a prerequisite, the whole collection of 102,765 documents was divided into two exclusive sets of training and testing with respective sizes of 71,936 and 30,829 documents. Then the BOW representation of documents is fed into the neural network. Output of TNG is used as a label for supervised training of the neural network so that it can learn the Topic Document Distribution as well as the Topic Word Distribution. Figure 9 shows the processing of our model.

The following stepwise implementation was implemented in our analysis.

1. To learn topics from the given dataset, TNG model was used to perform topic extraction, and to compute Topic Document Distribution θ as well as Topic Word Distribution ϕ.

2. As a standard, the optimal number of topics for a given dataset was selected by iteratively running the model for the number of topics ranging from 0 to 50. A trained model with finetuned parameters (hyper-parameters $\alpha,\ \beta, \gamma$ and δ) and suitable number of topics k are saved for further comparisons.

3. *tanh* activation function is used in our neural networks for hidden layers and *softmax* activation function is used at the output layer of neural networks. We used the stochastic gradient descent optimizer and categorical cross entropy as the loss function.

Compared with traditional Machine Learning, Deep Learning performs better when it comes to complex problems such as image classification, natural language processing and speech recognition. Therefore, we want to explore whether combined with deep learning can make the results of topic modeling better. Furthermore, LDA takes more times to process a large number of corpus. The proposed models are believed to reduce the computation cost required by LDA to extract topics (themes) of a huge corpus. We expect that running deep learning can improve accuracy and reduce training time.

In deep learning, adding more hidden layers can reduce the training errors. However, due to overfitting and high variance, adding too many layers will cause the result to have a high generalization error. Therefore, in this thesis, we only consider and compare 2 layers and 3 layers neural networks.

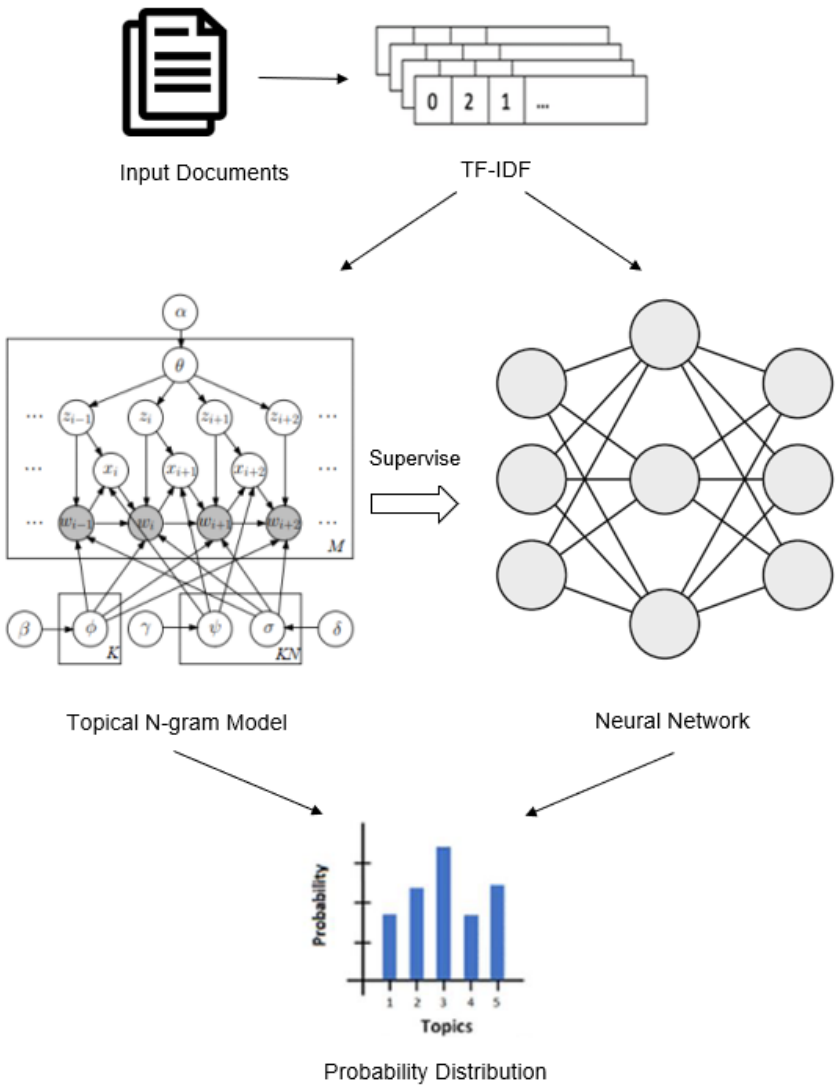

Figure 9: Processing of Deep Topical N-gram Model

Chapter 3 COVID-19 Data and Pre-processing

In this chapter, we describe the datasets to be analyzed and how we pre-process the data. In Section 3.1, we describe our dataset in detail and how we collect the data. Then we present how we clean the data before training in Section 3.2.

3.1 Data Description

The first dataset we use were published on Kaggle, coming from COVID-19 News Articles Open Research Dataset [29]. It has over 7,000 Canadian Broadcasting Corporation (CBC) news articles for the period of January 9, 2020 to May 3, 2020, with full text about COVID-19 and the coronavirus family of viruses. Table 2 shows examples of the raw data. On average, there are about 50 articles per day, with fewer articles in January of 2020 (about 400 in total) and the most articles in April of 2020 (about 2700 in total). This is a dataset collected by Web Scraping from CBC news search result regarding coronavirus [30].

Table 2: Raw Data of COVID- 19 CBC News

	authors	title	publish_date	description	text	url
0	[]	'More vital now:' Gay-straight alliances go vi...	2020-05-03 1:30	Lily Overacker and Laurell Pallot start each g...	Lily Overacker and Laurell Pallot start each g...	https://www.cbc.ca/news/canada/calgary/gay-str...
1	[]	Scientists aim to 'see' invisible transmission...	2020-05-02 8:00	Some researchers aim to learn more about how t...	This is an excerpt from Second Opinion, a week...	https://www.cbc.ca/news/technology/droplet-tra...
2	['The Canadian Press']	Coronavirus: What's happening in Canada and ar...	2020-05-02 11:28	Canada's chief public health officer struck an...	The latest: The lives behind the numbers: Wha...	https://www.cbc.ca/news/canada/coronavirus-cov...
3	[]	B.C. announces 26 new coronavirus cases, new c...	2020-05-02 18:45	B.C. provincial health officer Dr. Bonnie Henr...	B.C. provincial health officer Dr. Bonnie Henr...	https://www.cbc.ca/news/canada/british-columbi...
4	[]	B.C. announces 26 new coronavirus cases, new c...	2020-05-02 18:45	B.C. provincial health officer Dr. Bonnie Henr...	B.C. provincial health officer Dr. Bonnie Henr...	https://www.cbc.ca/news/canada/british-columbi...

Web Scraping is a technique employed to extract large amounts of data from websites which are saved as a local file in the computer or a database in the table (spreadsheet) format [37-39]. Data displayed on most websites can only be viewed using a web browser which often does not provide the function of saving data for the personal use. The only option is to manually copy and paste the data, which is a tedious task that may take hours or even days to complete. Web Scraping is a technology that automates the process and can complete the same task in a short period of time. The web scraping software automatically loads and extracts data from multiple pages of the website

according to users' requirements. It can be custom built for a specific website or be configured to be used for any website. With the push of a button, users can easily save the data available on the website to a file on their own computers. An illustrative diagram is presented in Figure 10.

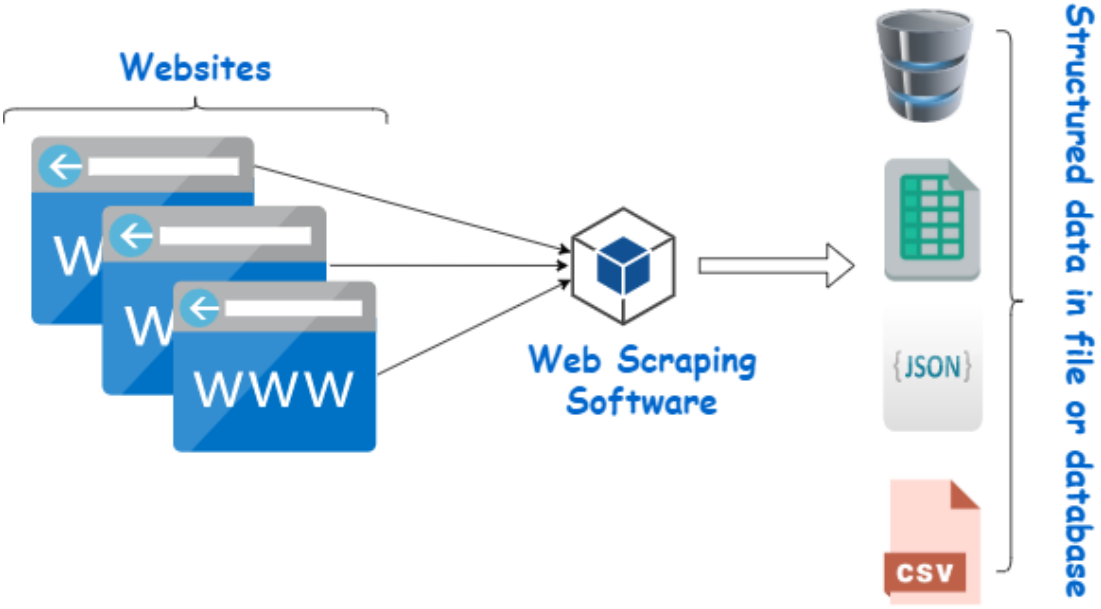

Figure 10: Web Scraping Process [63]

The second dataset comes from the COVID-19 open research data set (CORD-19) prepared by the White House and a coalition of leading research groups, such as Microsoft Research, IBM, the National Library of Medicine, etc. [40]. CORD-19 is a resource of over 200,000 scholarly articles collected from 2000 to 2020, including over 94,000 with full text, about COVID-19 SARS-CoV-2, or related coronaviruses. This freely available dataset provides to the research community an opportunity of using natural language processing and other AI techniques to generate new insights in response to the ongoing fight against COVID-19. There is a growing urgency for such types of research because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.

In this dataset, there are 14 variables, including CORD UID, SHA, paper title, paper authors, license, DOI, PMC ID, PubMed ID, source, abstract, publish time, journal,

JSON files (each paper is represented as a single JSON object), and URL. Table 3 presents a table of a sample of the data.

Table 3: Example of CORD-19 Raw Data

cord_uid	2b73a28n	02tnwd4m	ejv2xln0
sha	348055649b6b8cf2b9a376498df9bf41f7123605	6b0567729c2143a66d737eb0a2f63f2dce2e5a7d	06ced00a5fc04215949aa72528f2eeaae1d58927
source_x	PMC	PMC	PMC
title	Role of endothelin-1 in lung disease	Nitric oxide: a pro-inflammatory mediator in lung disease?	Surfactant protein-D and pulmonary host defense
doi	10.1186/rr44	10.1186/rr14	10.1186/rr19
pmcid	PMC59574	PMC59543	PMC59549
pubmed_id	11686871	11667967	11667972
license	no-cc	no-cc	no-cc
abstract	Endothelin-1 (ET-1) is a 21 amino acid peptide with diverse biological activity that has been implict...	Inflammatory diseases of the respiratory tract are commonly associated with elevated production of …	Surfactant protein-D (SP-D) participates in the innate response to inhaled microorganisms and organic...
publish_time	2001-02-22	2000-08-15	2000-08-25

authors	Fagan, Karen A; McMurtry, Ivan F; Rodman, David M	Vliet, Albert van der; Eiserich, Jason P; Cross, Carroll E	Crouch, Erika C
journal	Respir Res	Respir Res	Respir Res
pdf_json_files	document_parses/pdf_json/348055649b6b8cf2b9a376498df9bf41f7123605.json	document_parses/pdf_json/6b0567729c2143a66d737eb0a2f63f2dce2e5a7d.json	document_parses/pdf_json/06ced00a5fc04215949aa72528f2eeaae1d58927.json
pmc_json_files	document_parses/pmc_json/PMC59574.xml.json	document_parses/pmc_json/PMC59543.xml.json	document_parses/pmc_json/PMC59549.xml.json
url	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC59574/	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC59543/	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC59549/

```
"COVID" OR "COVID-19" OR
"Coronavirus" OR "Corona virus"
OR "2019-nCoV" OR "SARS-CoV"
OR "MERS-CoV" OR "Severe Acute
Respiratory Syndrome" OR "Middle
East Respiratory Syndrome"
```

Figure 11: CORD-19 Paper Query

All papers are retrieved based on the queries indicated in Figure 11. Figure 12 shows the distribution of papers per year in CORD-19. There is a spike in publications occurs in 2020 in response to COVID-19. Over 47,000 papers and 7,000 preprints on COVID-19 and coronaviruses have been released since the beginning of 2020, comprising nearly

40% of papers in the dataset. The dataset consists predominantly of papers in Medicine (55%), Biology (31%), and Chemistry (3%), which together constitute almost 90% of the corpus.

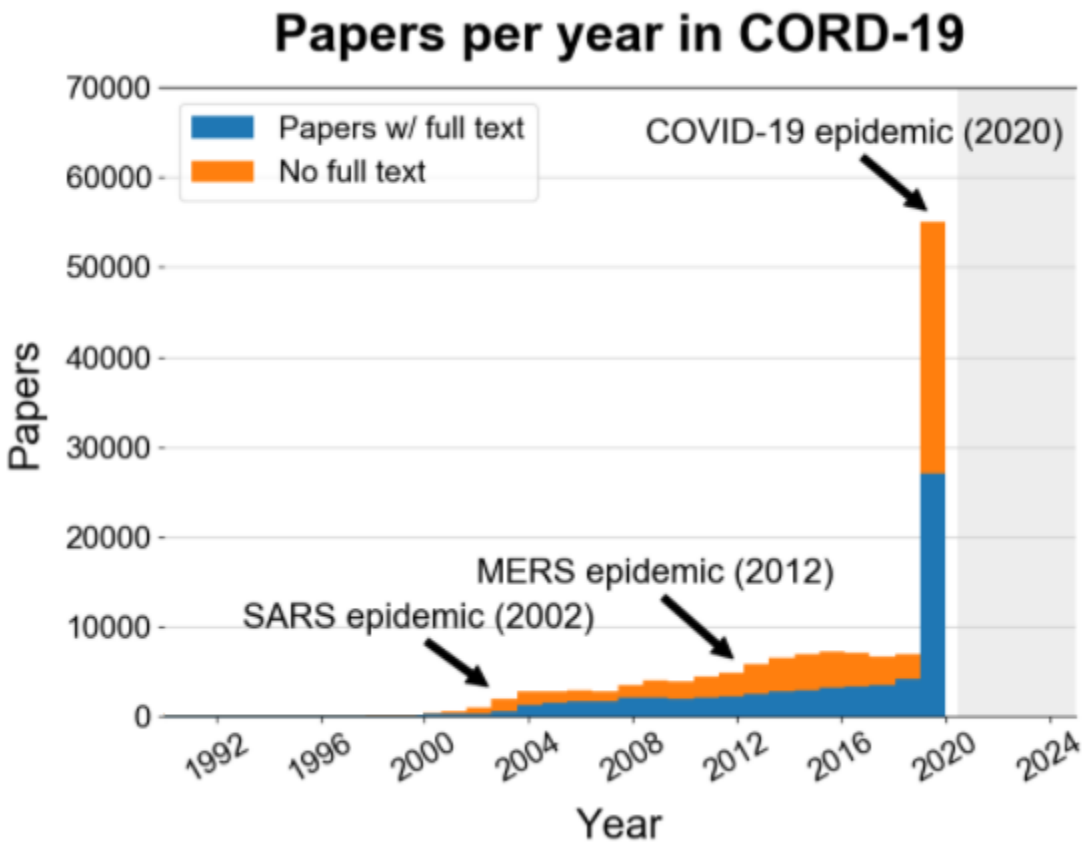

Figure 12: The Distribution of Papers in CORD-19 by Year in CORD-19. A spike in publications occurs in 2020 in response to COVID-19 [40]

3.2 Data Pre-processing

In Natural Language Processing approaches, it is essential to clean the text data before training. The primary purpose of the preprocessing steps is to reduce the number of unimportant words without distorting the general meaning. Here we are mainly interested in studying how the content influences the number of reviews and rating. The preprocessing is as follows:

- Remove all URL links and authors. The URLs and authors do not contain the sentiment information, thus, being deleted from the dataset by using Pandas library in Python [32].

- Remove the punctuation and whitespace. By applying word2Vec algorithm, punctuation and whitespace will be treated as words when using trained embeddings as the input to a sequence model for a part of speech tagging task or sentiment classification task. Python offers functions to remove these unimportant characters.

- Remove all non-ASCII, non-English characters, and stop words. Stop words usually refer to the most common words in a language, such as “the”, “an”, and “than”. The classic method is based on removing the stop words obtained from precompiled lists. We remove the words using the NLTK Library in Python.

- Drop missing (NaN) and duplicated values. Duplicate elimination algorithms will be applied in this step. The method of eliminating duplicate records in a file has been the following: first, the file is sorted using an external merge-sort in order to bring all duplicate records together; then, a sequential pass is made through the tile comparing adjacent records and eliminating all but one of the duplicates.

- Tokenize articles to sentences, then sentences to words by using the SpaCy [31]. SpaCy is a free open-source library for Natural Language Processing. First, the raw text is split on whitespace characters, similar to text.split (' '). Then, the tokenizer processes the text from left to right.

Chapter 4 Case Study 1 – COVID-19 News Articles Open Research Dataset Analysis

In this chapter, we apply three standard models, TD-IDF model, and N-gram model in contrast to our proposed model (described in Section 2.11) to analyze the COVID-19 news articles described in Section 3.1. Sections 4.1 and 4.2 show the results obtained from the three standard models. Section 4.3 reports the results obtained from LDA model and our proposed model together with the visualization in word cloud. The findings are summarized in Section 4.5

4.1 TF-IDF Model

To reveal possible patterns, we divide the data into one word (unigram), two words (bigrams) and three words (trigrams), respectively, and calculate the corresponding TF-IDF scores.

Figure 13 shows the unigram, suggesting that "Wuhan", "pandemic", "province", "physical distancing", and "care" are very common in the news.

Figure 14 presents the bigrams, indicating that "deaths Jan", “social distancing", "term care", and "physical distancing" are hot topics in the news.

Figure 15 displays the trigrams, revealing that "Beijing railway station", “Diamond Princess Cruise", "term care homes", and "chief medical officer" are top topics in the news.

With bigrams and trigrams, we gain a better understanding on the Canadian news focus. When the epidemic first broke out in January of 2020, news mainly focused on reporting the situation in Wuhan, China with the risk of COVID-19 in Canada regarded low. In February of 2020, more news about the Wuhan epidemic and how the Chinese government took action on controlling the spread of the virus. Also, outbreaks occurred on several cruise ships. Diamond Princess cruise is the most severely affected cruise ship, and tourists had to be quarantined on the cruise ship. The COVID-19 pandemic began to break out in Canada in March of 2020. It then became the main topic of the news. Public

health officers and the Prime Minister Justin Trudeau had a lot of announcements and they suggested people to follow social distancing to reduce the virus transmission. Long-term care homes had been the epicenter of the COVID-19 pandemic in Canada since April of 2020, and public health officers wanted people to practise physical distancing. May of 2020 became the peak period of CVOID-19 and, physical distancing was still mandatory for people to follow.

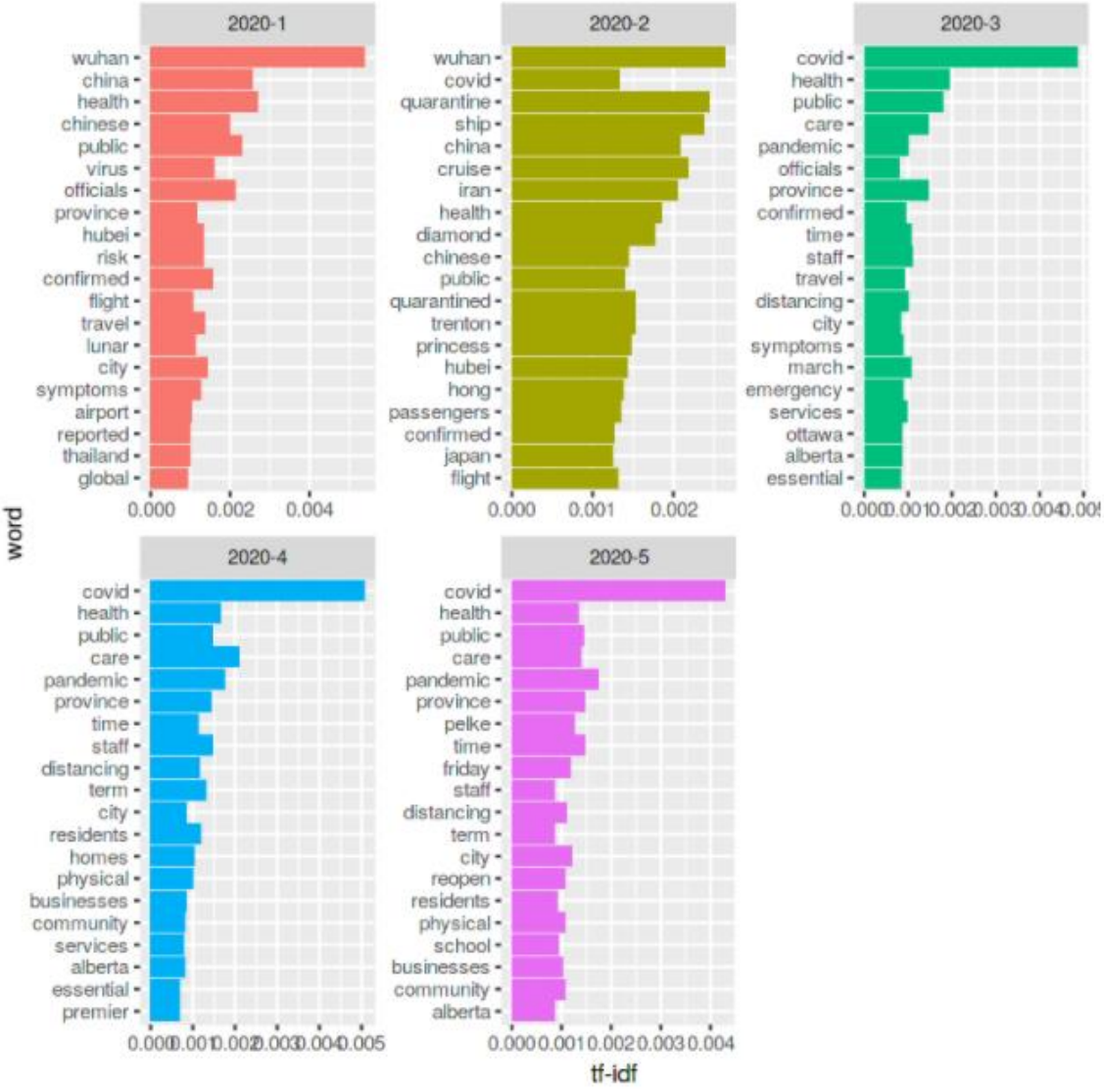

Figure 13:TF-IDF with Unigram

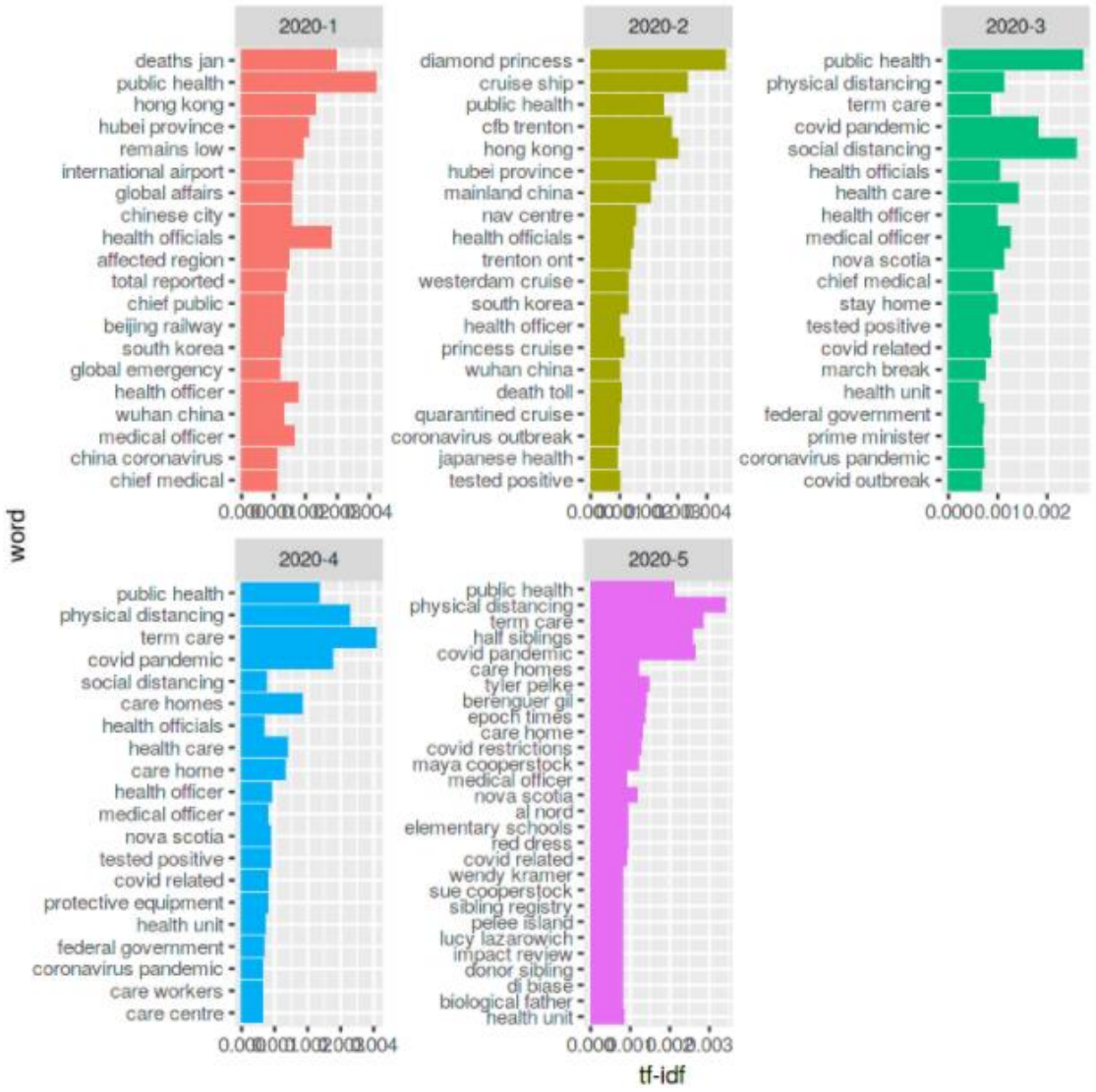

Figure 14: TF-IDF with Bigram

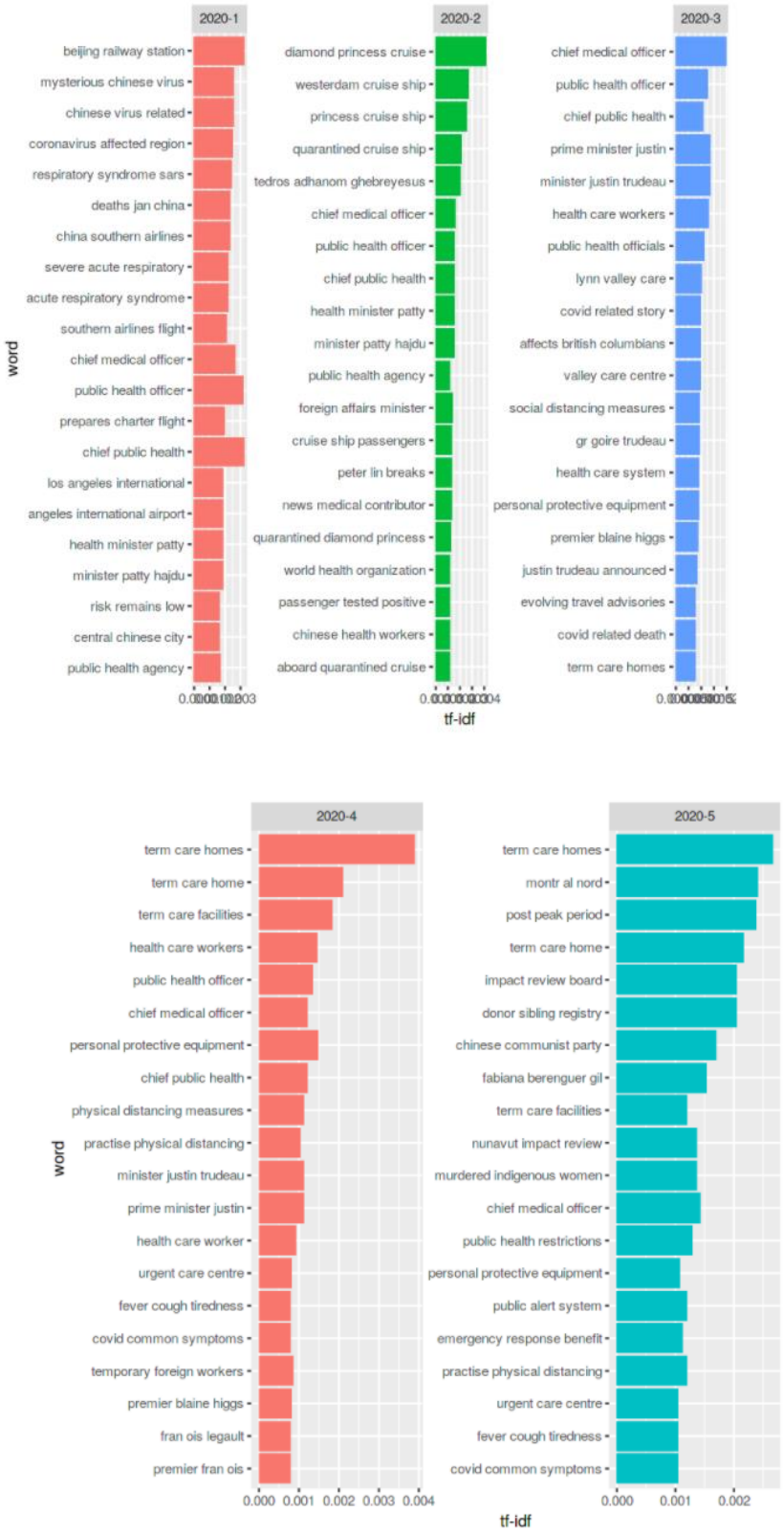

Figure 15: TF-IDF with Trigram

Furthermore, we calculate the frequency of individual words and repot in Tables 4-6 the top ten words with their frequency and relative frequency with the three high-frequency words, "Pandemic", "Coronavirus", and "Health".

Table 4 shows that "Pandemic COVID" appears 198 times in the news, and "Pandemic Canada", "Pandemic Care" and "Pandemic Response" show at least 100 times. Some provinces and cities (e.g., Edmonton, Alberta, Manitoba, Calgary) combine with "Pandemic" at least 50 times.

Table 5 illustrates that “Coronavirus Outbreak” has been mentioned 114 times in the news dataset. The words with countries or areas, for example, “Coronavirus China”, “Coronavirus Canada”, and “Coronavirus World”, display at least 50 times.

Table 6 demonstrates that “COVID Health” appears 243 times in the CBC news dataset, and it shows together with “Health Care” 110 times. The words, “Health Chief”, “Health Officials”, and “Health Workers” show at least 50 times.

Table 4: The top 10 frequency words associated with "Pandemic"

Word 1	**Word 2**	**Times (n) (Proportion)**
Pandemic	COVID	198 (5.27%)
Coronavirus	Pandemic	140 (1.06%)
Pandemic	Canada	138 (1.01%)
Pandemic	Care	134 (0.90%)
Pandemic	Response	127 (0.71%)
Pandemic	Health	126 (0.69%)
Pandemic	Workers	126 (0.69%)
Pandemic	Calgary	69 (0.51%)
Pandemic	Experts	62 (0.48%)
Pandemic	Manitoba	57 (0.45%)

Table 5: The top 10 frequency words associated with "Coronavirus"

Word 1	**Word 2**	**Times (n) (Proportion)**
Coronavirus	Outbreak	114 (2.06%)
Coronavirus	China	91 (1.65%)
Coronavirus	Canada	87 (1.58%)
Coronavirus	Health	63 (1.14%)
Coronavirus	World	56 (1.01%)
Coronavirus	Fears	55 (1.00%)
Coronavirus	April	49 (0.89%)
Coronavirus	Happening	46 (0.83%)
Coronavirus	Spread	41 (0.74%)

Table 6: The top 10 frequency words associated with "Health"

Word 1	**Word 2**	**Times (n) (Proportion)**
COVID	Health	243 (7.23%)
Health	Care	110 (3.27%)
Health	Public	84 (2.50%)
Health	Officials	76 (2.26%)
Health	Workers	72 (2.14%)
Coronavirus	Health	63 (1.88%)
Health	Officer	53 (1.58%)
Health	Unit	41 (1.22%)
Health	Chief	34 (1.01%)
Health	Thunder	34 (1.01%)

4.2 N-Gram Model

We are interested in visualizing how words may appear together. We particularly examine the structures of two-words (i.e., Bigram) and of three-words (i.e.: Trigram), and display them in Table 7 and Table 8, respectively. From the results, we can clearly see that the frequency of words about “Public Health” is the most, with a total of 7,545 times. Public health has a high degree of attention. However, compared with TF-IDF, the word "Public Health" appears more frequently in the *N*-gram model. This is because *N*-gram does not consider the weight of phrases in a sentence, but only considers whether they are linked together.

Table 7: The top 10 frequency words of the Bigram Model

Word 1	**Word 2**	**Times (n) (Proportion)**
Public	Health	7545 (0.85%)
Health	Care	3932 (0.44%)
Term	Care	3136 (0.35%)
Physical	Distancing	2770 (0.31%)
COVID	Pandemic	2699 (0.30%)
Health	Officials	2670 (0.30%)
Tested	Positive	2489 (0.28%)
CBC	News	2464 (0.28%)
Federal	Government	1957 (0.22%)
Medical	Officer	1779 (0.20%)

Table 8: The top 10 frequency words of the Trigram Model

Word 1	**Word 2**	**Word 3**	**Times (n) (Proportion)**
Health	Care	Worker	1199 (0.44%)
Term	Care	Homes	1159 (0.33%)
Chief	Medical	Officer	1135 (0.31%)
Public	Health	Officials	1104 (0.29%)
Public	Health	Officer	989 (0.27%)
Chief	Public	Health	892 (0.24%)
Personal	Protective	Equipment	870 (0.24%)
Minister	Justin	Trudeau	845 (0.23%)
Prime	Minister	Justin	844 (0.23%)
World	Health	Organization	835 (0.23%)

4.3 Topic Modeling and Visualization

Topic Coherence [24] is a measure used to evaluate topic models and determine the optimal number of topics. The coherence of a topic measures the degree of semantic similarity among high scored words. The measure helps distinguish between topics that are semantically interpretable and topics that are artifacts of statistical inference. The steps of computing topic coherence of a topic model are as follows:

1. Select the top n frequently occurring words in each topic, where n is a positive integer.

2. Compute pairwise scores for each of the words selected above and aggregate all the pairwise scores to calculate the coherence score for a particular topic:

$$Coherence = \sum_{i<j} score(w_i, w_j)$$

   Where top n words, $w_1, \dots w_n$, are used to describe the topic, and score (w_i, w_j) is calculated by the probability of seeing both w_i and w_j co-occurring in a random document.

3. Take the mean of the coherence scores for all topics in the model to arrive at a score for the topic model.

4.3.1 Grid Search of the Optimal Number of Topics

A grid search algorithm is used to find the optimal number of topics in the CBC news. Figure 16 plots the LDA coherence score versus the number of the topics. The grid program shows that when the number of topic is 26, the coherence score is the largest (0.321), and the second largest value is achieved with 14 topics, with a score of 0.310. Here, the blue line connects the coherence scores against the number of topics, where the vertical axis represents the coherence score, and the horizontal axis stands for the number of topics. This figure is constructed by the *genism* library.

Next, we evaluate the quality of the topic modeling to select the optimal of number of topics. We use K as 14 and 26 respectively to draw the *PyLDAvis* plots of the LDA model for comparison. *PyLDAvis* is created based on *LDAvis*, a web-based interactive visualization of topics estimated using the Latent Dirichlet Allocation. More specifically, *PyLDAvis* is a Python library for the interactive topic model visualization which helps users interpret the topics. *PyLDAvis* provides two visualization panels as shown in Figure 17, each bubble in the figure represents a topic. The larger the bubble is, the more common the topic is. A good topic model will display quite large non-overlapping bubbles throughout the chart, rather than being clustered in one quadrant. Models with too many topics usually have a lot of overlap, and small-sized bubbles are concentrated in one area of the chart. The words and bars on the right represent the most salient term in the corpus. When $K = 14$, *PyLDAvis* bubbles are more evenly distributed in the quadrant with less overlap, which shows that 14 is the optimal number of topics we are looking for.

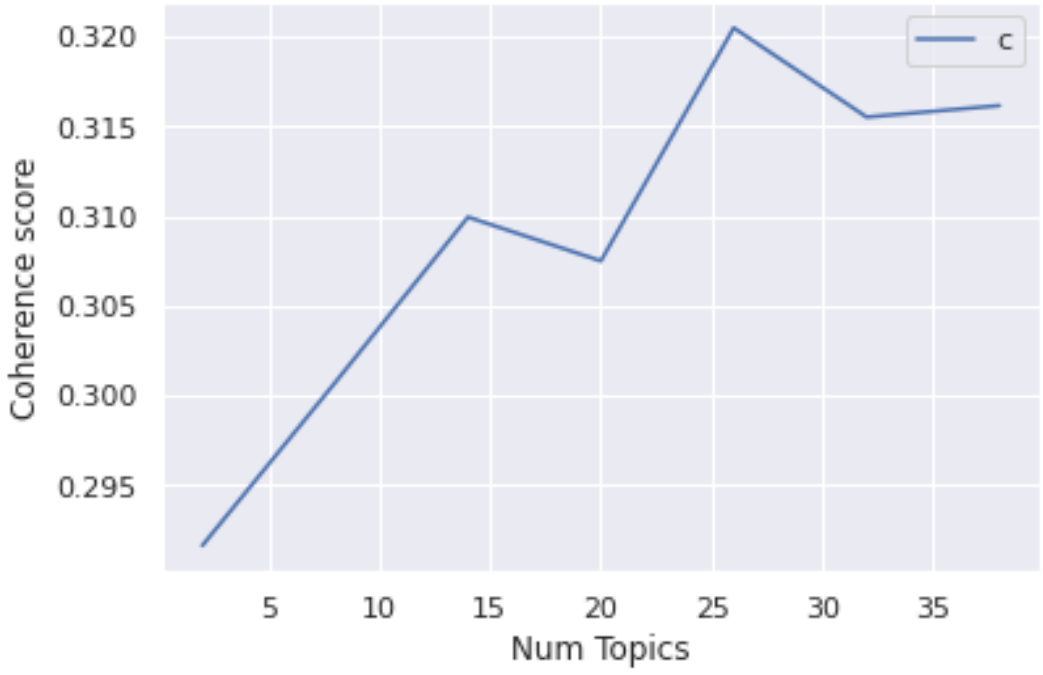

Figure 16: LDA topic coherence score versus the number of topics

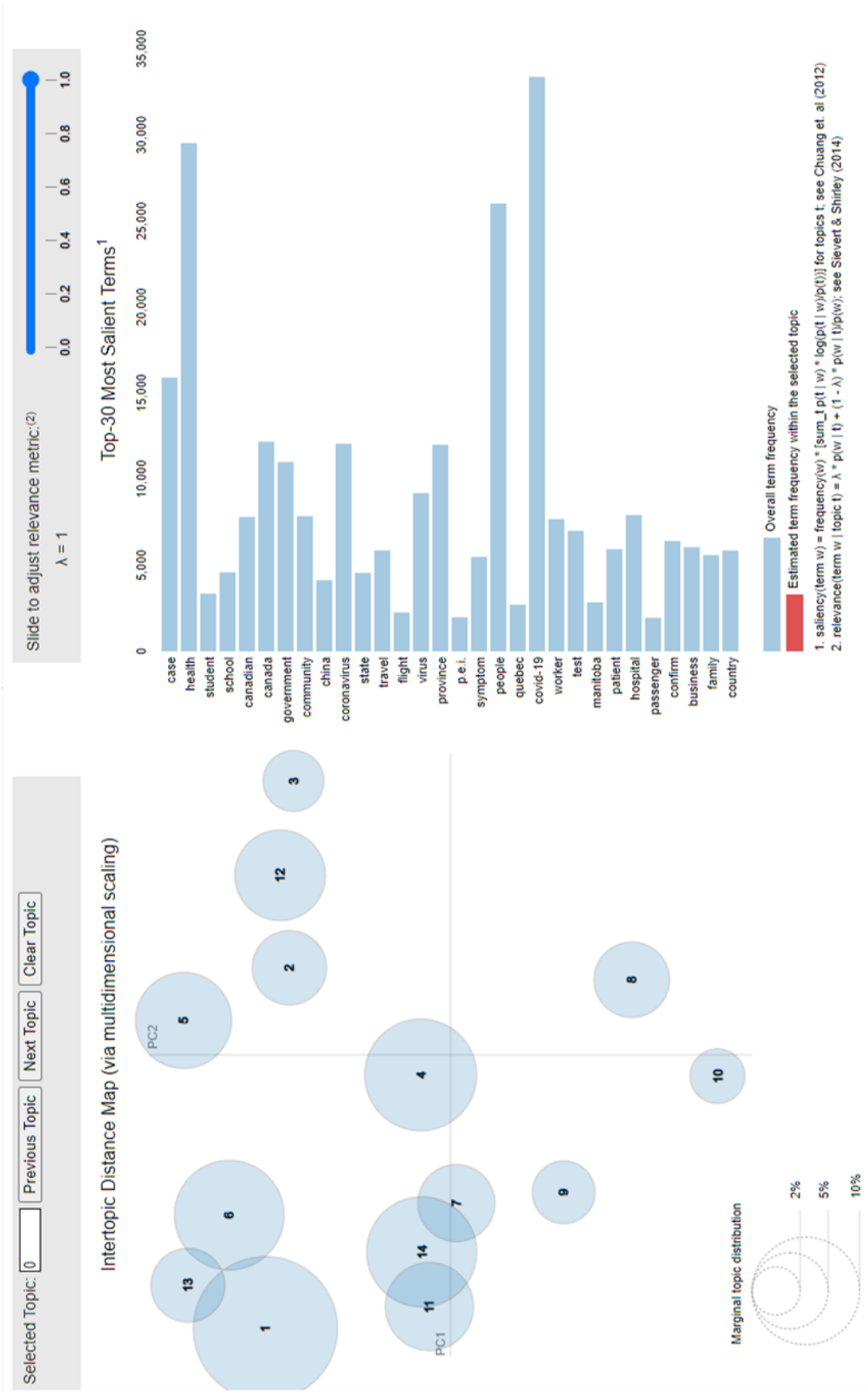

Figure 17: Top-30 Most Salient Terms

4.3.2 LDA Visualization with Word Cloud

To visualize each topic, we use a word cloud to display top 15 words in each topic. The word cloud is a visualization tool to display words in a text according to the frequency the words appear in the text. The size of each word indicates its importance, or the frequency; the larger the word size, the more frequent the word in a topic. In addition, the words that dominate the word cloud is most likely directly related to the topic.

- Topic #0 : Business

Topic #0 talks about business. There are some dominated words in Figure 18 that represent the topic, such as “restaurant”, “business”, “customer”. Those words can be summarized as the impact of COVID-19 on businesses and government response. Also, the words “program”, “small”, “support” and “government” indicate that government has programs to support small business.

Figure 18: Word Cloud Topic #0

- Topic #1: Tested and confirmed cases

Figure 19 shows the word cloud of topic #1. We find that the word symptom, health and test as the most dominant. Other words like “new”, “number”, “case” and “province” indicate new cases in each province. From those dominant words, it can be concluded that the topic is about COVID-19 confirmed cases in different provinces.

Figure 19: Word Cloud Topic # 1

- Topic #2: Government of Canada and Prime Minister

The word cloud visualization of topic #2 is shown by Figure 20. The words “Justin”, “CBC”, “news” and “border” are dominating in this topic segment. In addition, there are also listed words related to government such as “minister”, “federal”, and “prime”. According to those words, it can be concluded that topic #2 talks about Government of Canada and Prime Minister.

Figure 20: Word Cloud Topic #2

Topic #3: Outbreak and death in long-term care home

The topic discussed in topic #3 is clearly about the COVID-19 outbreak and death in long-term care homes, which can be seen from the terms “outbreak”, “death”, “patient”, “staff”, “hospital” and “long-term”. The word cloud of topic #3 can be seen in Figure 21.

Figure 21: Word Cloud Topic #3

- Topic #4: Prince Edward Island

According to the word cloud in Figure 22, the most appropriate topic to topic #4 is about the influence in Prince Edward Island (PEI). The word “Morrison” is related to Dr. Morrison who is the Chief Public Health Officer in PEI. Also, the words “fisherman”, “island” and “ferry” are related to PEI.

Figure 22: Word Cloud Topic #4

- Topic #5: Masks

Figure 23 shows the word cloud of topic #5. From the terms in the word cloud, it can be concluded that the corresponding topic of topic #5 is about wearing mask, which can be seen from the terms “mask”, “spread”, “advice” and “cough”. Topic #5 suggests that wearing mask is mandatory and can prevent the spread of the virus.

Figure 23: Word Cloud Topic #5

- Topic #6: Life during Lockdown

The words in topic #6 deal with people’s lives during the COVID-19 pandemic which can be seen from three dominant words in the word cloud, “family”, “food” and “friend”. The word cloud of topic #6 is shown by Figure 24.

Figure 24: Word Cloud Topic #6

- Topic #7: Cases in Canada

Topic #7 relates to the number of cases for the COVID-19 outbreak in Canada, from Figure 25, which can be seen from three dominant words in its word cloud, “outbreak”, “number” and “Canada”.

Figure 25: Word Cloud Topic #7

- Topic #8: Students

Topic #8, in Figure 26, talks about school and students during the COVID-19 pandemic. Especially, the words “school”, “student”, “university”, “teacher”, “child”, and “parent” are characters related to the topic.

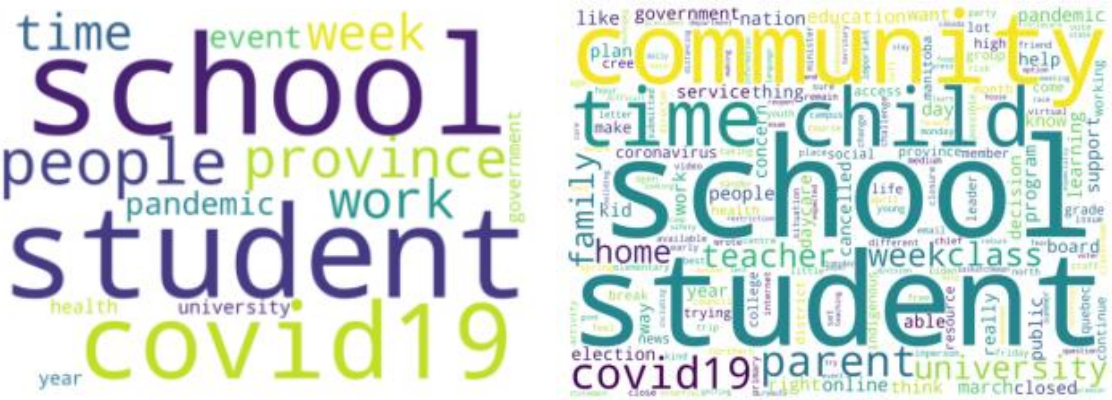

Figure 26: Word Cloud Topic #8

- Topic #9: Traveling

According to the word cloud in Figure 27, it is obvious that topic #9 is about traveling during the COVID-19 pandemic. The words “passenger”, “travel”, “flight” indicate the topic about traveling. Moreover, other words such as “quarantine”, “return” and “trip” are relevant to traveling.

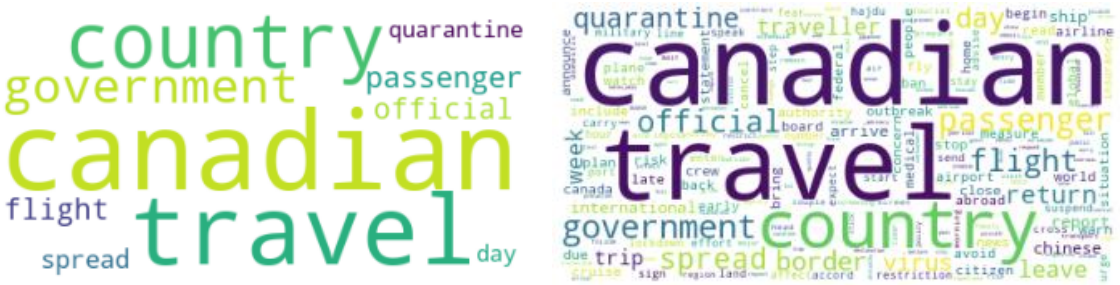

Figure 27: Word Cloud Topic #9

- Topic #10: Cases in other provinces

Topic #10, shown in Figure 28 is related to the Nova Scotia, Winnipeg, and some towns, in other provinces that had cases.

Figure 28: Word Cloud Topic #10

- Topic #11: US president election

The words in topic #11 concern the US president election during the COVID-19 pandemic which can be seen from three dominant words in the word cloud, “trump”, “election” and “president”. The word cloud of topic #11 is shown by Figure 29.

Figure 29: Word Cloud Topic #11

- Topic #12: Economy

Topic #12 shows that, as the coronavirus becomes serious, the Canadian economy is affected. Especially for the oil price was getting lower than usually. From Figure 30, we can find the words “oil”, “economy”, “crisis” and “price”.

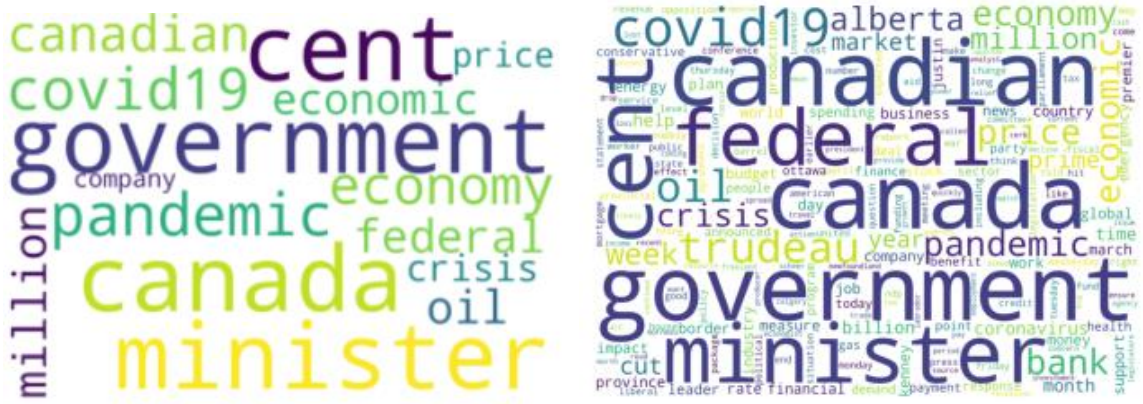

Figure 30: Word Cloud Topic #12

- Topic #13: Cancellation

Topic #13 talks about, with COVID-19 coming, many activities had to be cancelled. Figure 31 lists the words, “plan”, “event”, “schedule”, “cancel”, which indicate the topic about cancellation.

Figure 31: Word Cloud Topic #13

4.4 Findings and Discussions

In this chapter, we apply three models: TF-IDF model, LDA model, and our proposed model to analyze the COVID-19 news article open research datasets. We obtain some results. This dataset can be divided into 14 topics, each topic talking about the problems people are concerned about since January of 2020 to May of 2020. For example, “traveling”, “life during lockdown” and “masks” are the topics that we often hear from the news. Compared to the standard models, our proposed model produces two words results, which are more informative to understand. For example, “oil” with “oil price”, “tested” and “tested positive” and “health” and “health officer” are a lot easier to comprehend than a single word display with “oil”, “test” and “health”. From these results, we conclude that the focus of news is more on the impact of COVID-19 on people’s daily lives. For example, people have to wear masks to go out, millions of companies face survival threats, and the inconvenience caused by rising prices. Moreover, the government has played a key role in dealing with the pandemic. Also, in the TF-IDF approach, words that are more common in one sentence but less common in others may be given more weights.

Chapter 5 Case Study 2 - COVID-19 Open Research Dataset Analysis

In this chapter, we apply the same methods considered in Chapter 4 to analyze the COVID-19 open research dataset. Sections 5.1 and 5.2 show the results obtained from the TF-IDF model and the N-gram, respectively. Section 5.3 reports the result of the LDA model and our proposed model together with the visualization display. The findings are summarized in Section 5.4.

5.1 TF-IDF Base Model

To indicate possible patterns, we divide the data into one word (unigram), two words (bigrams) and three words (trigrams), respectively, and then calculate the corresponding TF-IDF scores.

- Figures 32 and 33 show the unigram, suggesting that "CI", "chest", "surgery", "ncov", and "igg" are common in the COVID-19 papers.

- Figures 34 and 35 present the bigrams from January of 2020 to September of 2020. We find that "coronavirus disease", "disease COVID", "COVID infection", and "coronavirus pneumonia " are high frequency words in the papers.

- Figures 36 to 39 display the trigrams, showing that "coronavirus disease COVID", "severe acute respiratory", "acute respiratory syndrome" and "coronavirus ncov spike" are top words in the papers.

With bigrams and trigrams, we get a better understanding of the key features of COVID-19-related papers. When the epidemic first broke out in January of 2020, the articles focused on the comparison of COVID-19 to the related diseases, such as "respiratory syndrome" and "pneumonia". There were also some articles explaining how COVID-19 was contagious. In February of 2020, more articles studied the symptoms of COVID-19. Some articles used regression models to predict the trend of the epidemic to gain understanding of the disease development. The COVID-19 pandemic has spread across the world since March of 2020. Almost all articles focused on COVID-19 and its daily

reports. Since April of 2020, articles mainly studied how the virus spreads, proposed some systems or models on topics such as speed control system, health protection guideline and unseen enemy mobilizing. In May of 2020, the world was still locked down. The UK proposed the herd immunity for the first time. Researchers published articles explaining whether the herd immunity was reliable. Researchers made new findings related to COVID-19 from June of 2020 to August of 2020, such as “bacterial foraging optimization”, “multisystem inflammatory syndrome”, and “angiotensin converting enzyme”.

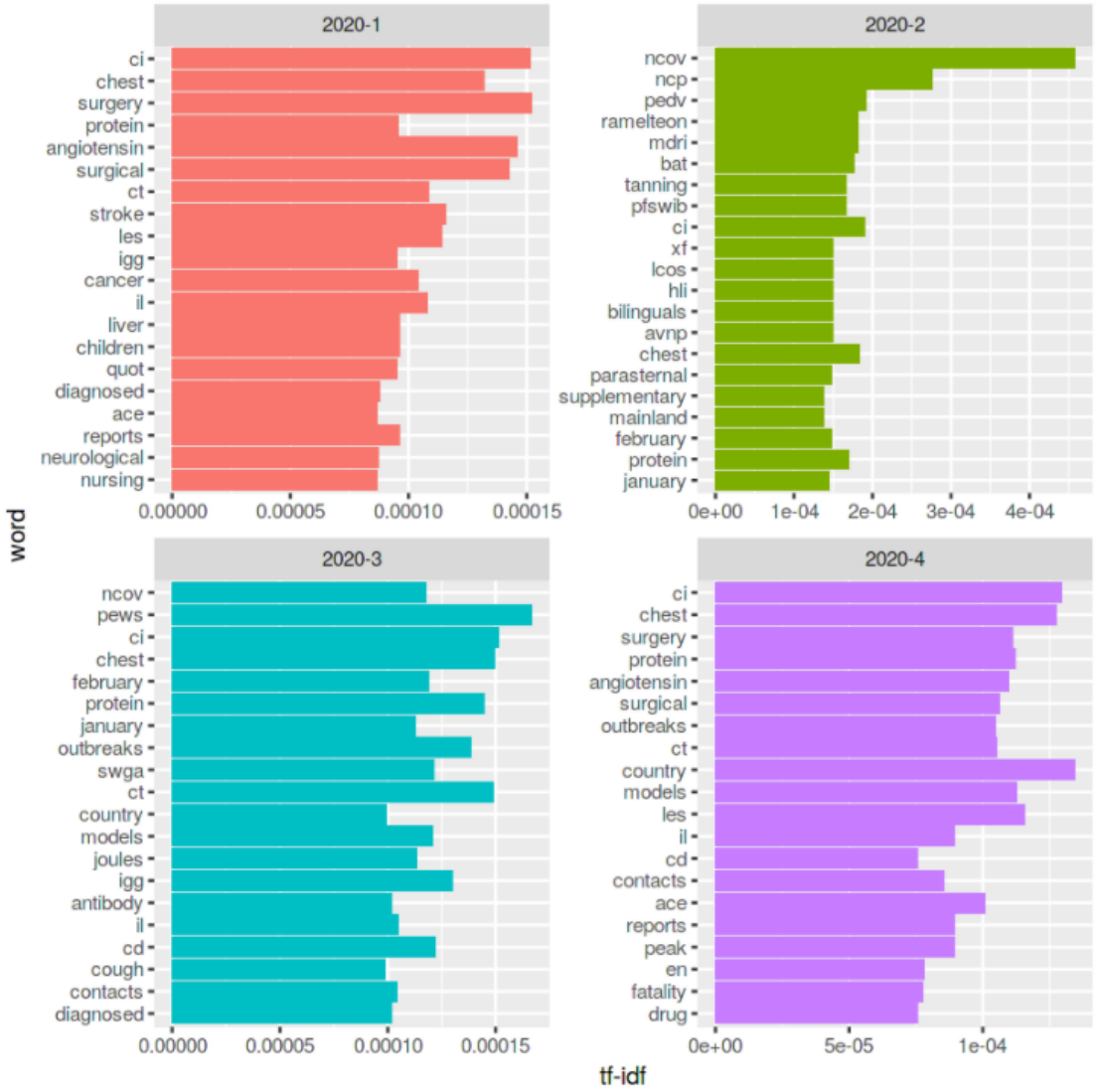

Figure 32: Unigram from January to May, 2020

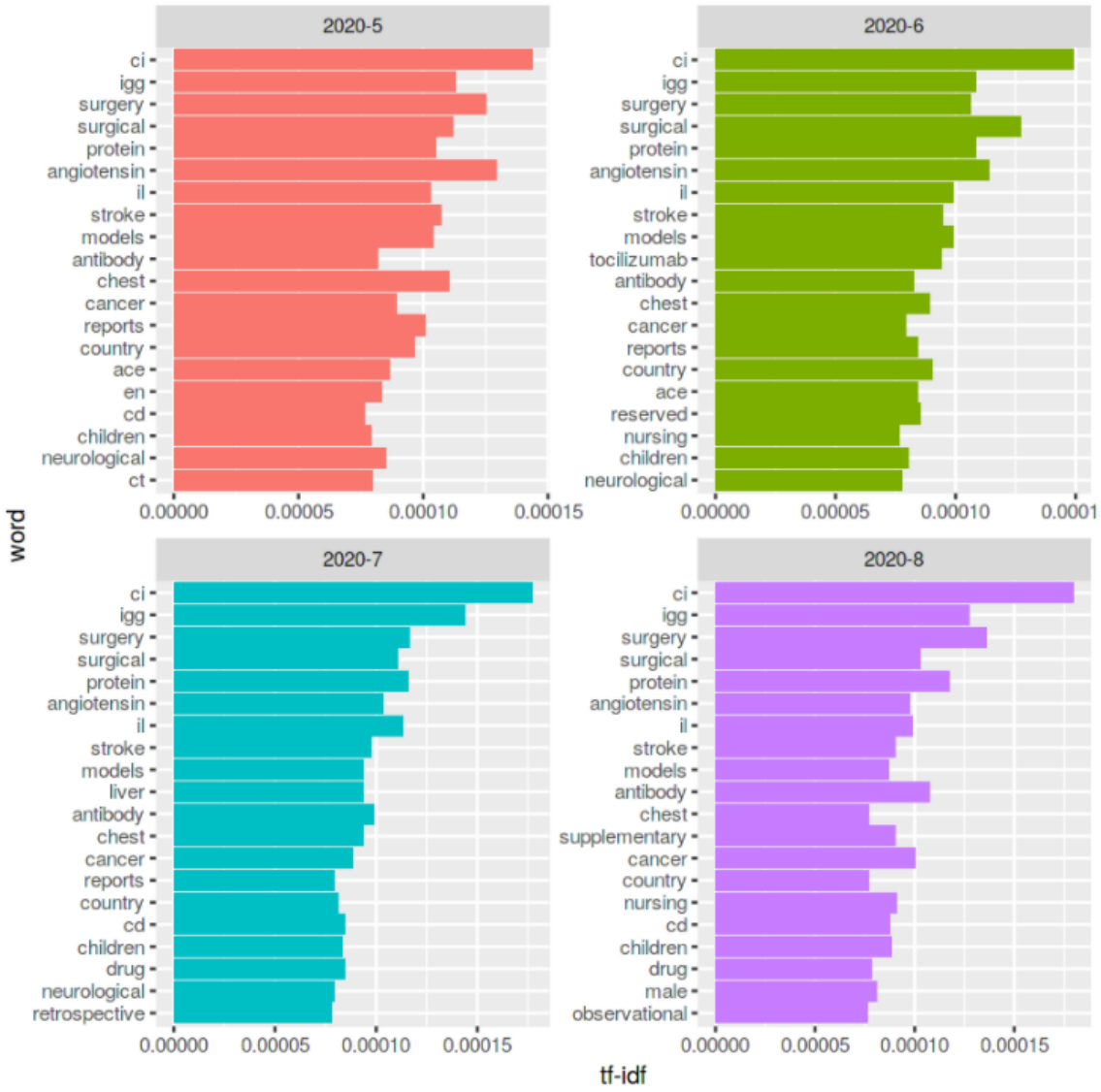

Figure 33: Unigram from June to August, 2020

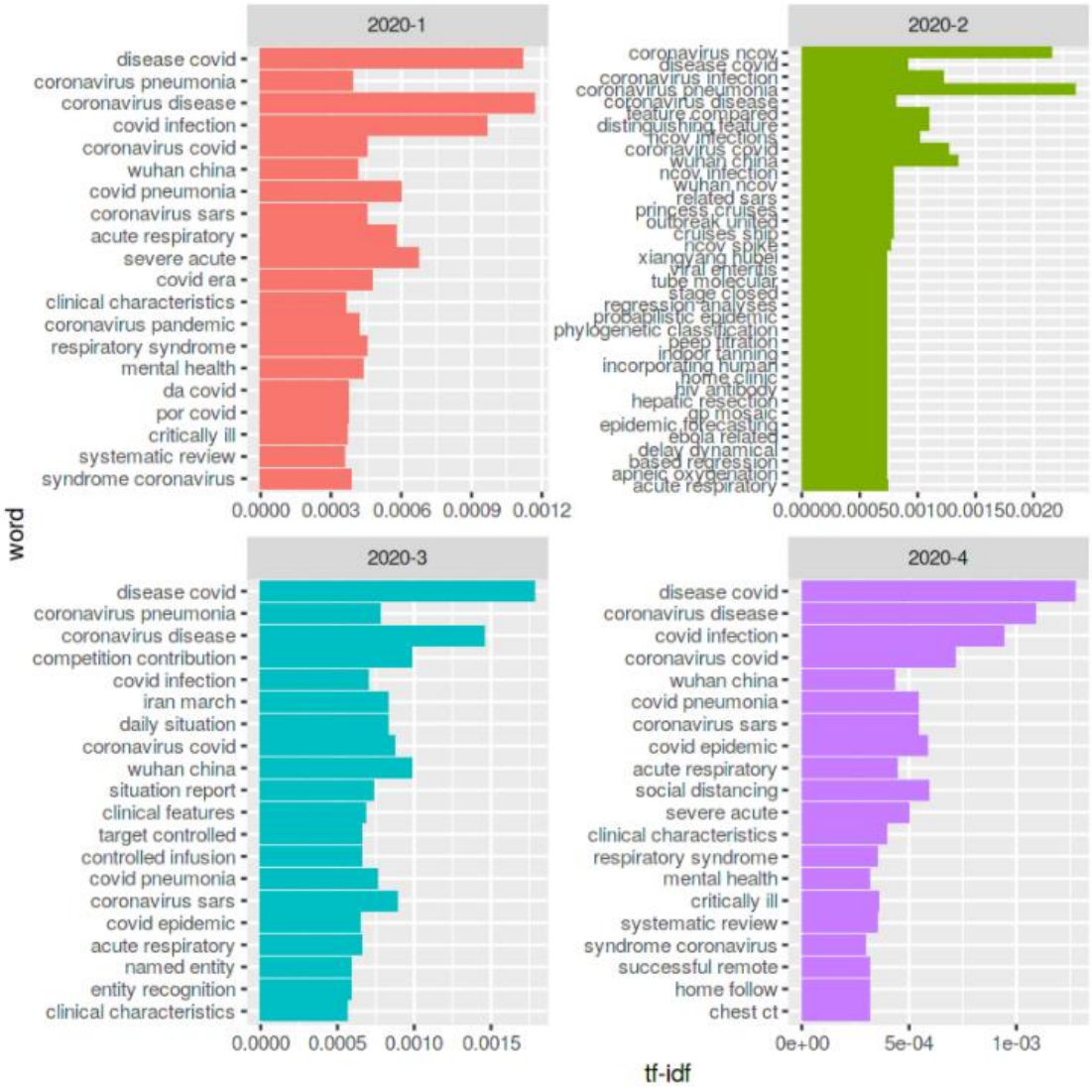

Figure 34: Bigram from January to May, 2020

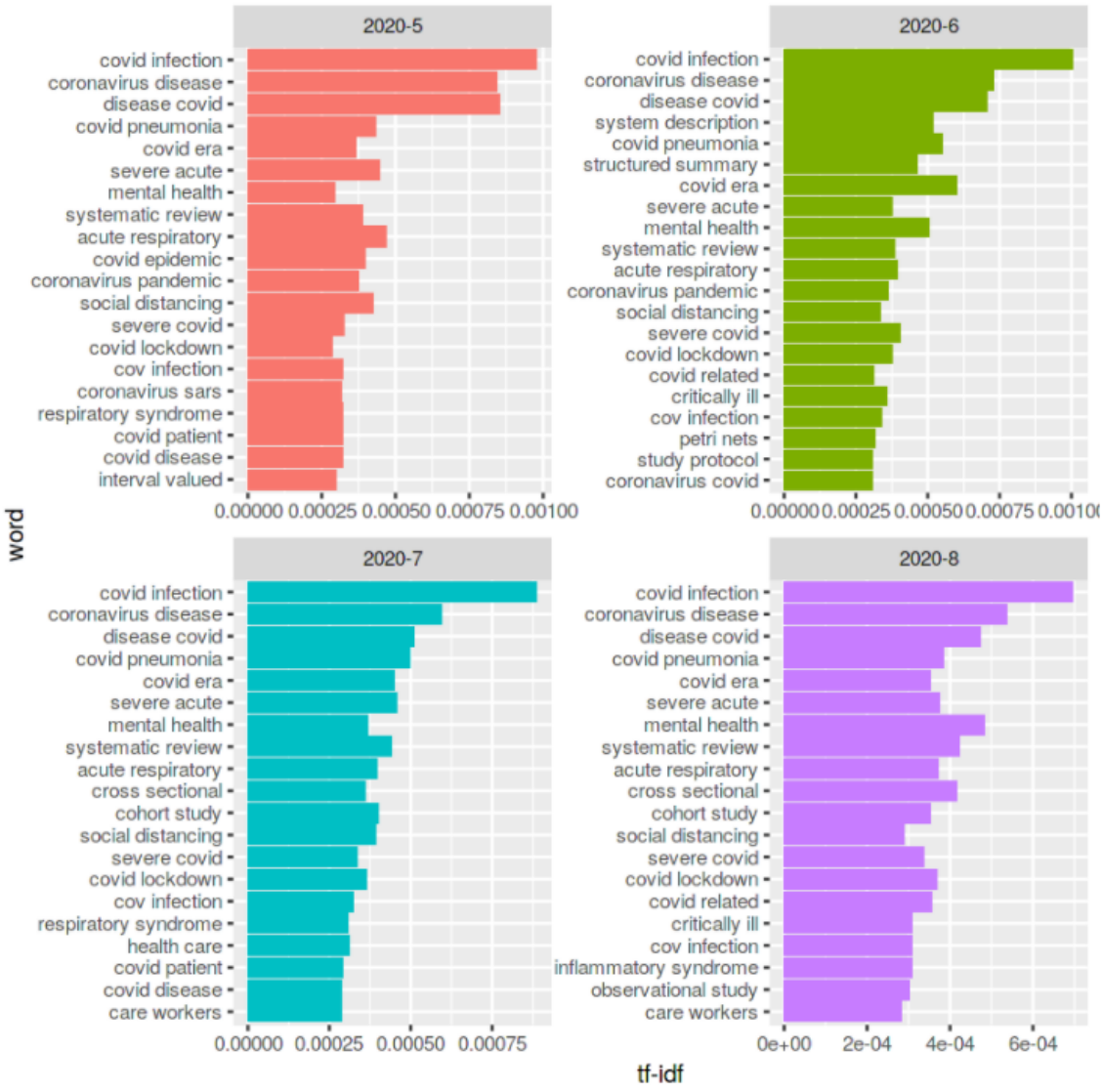

Figure 35: Bigram from June to August, 2020

Figure 36: Trigram for January and February, 2020

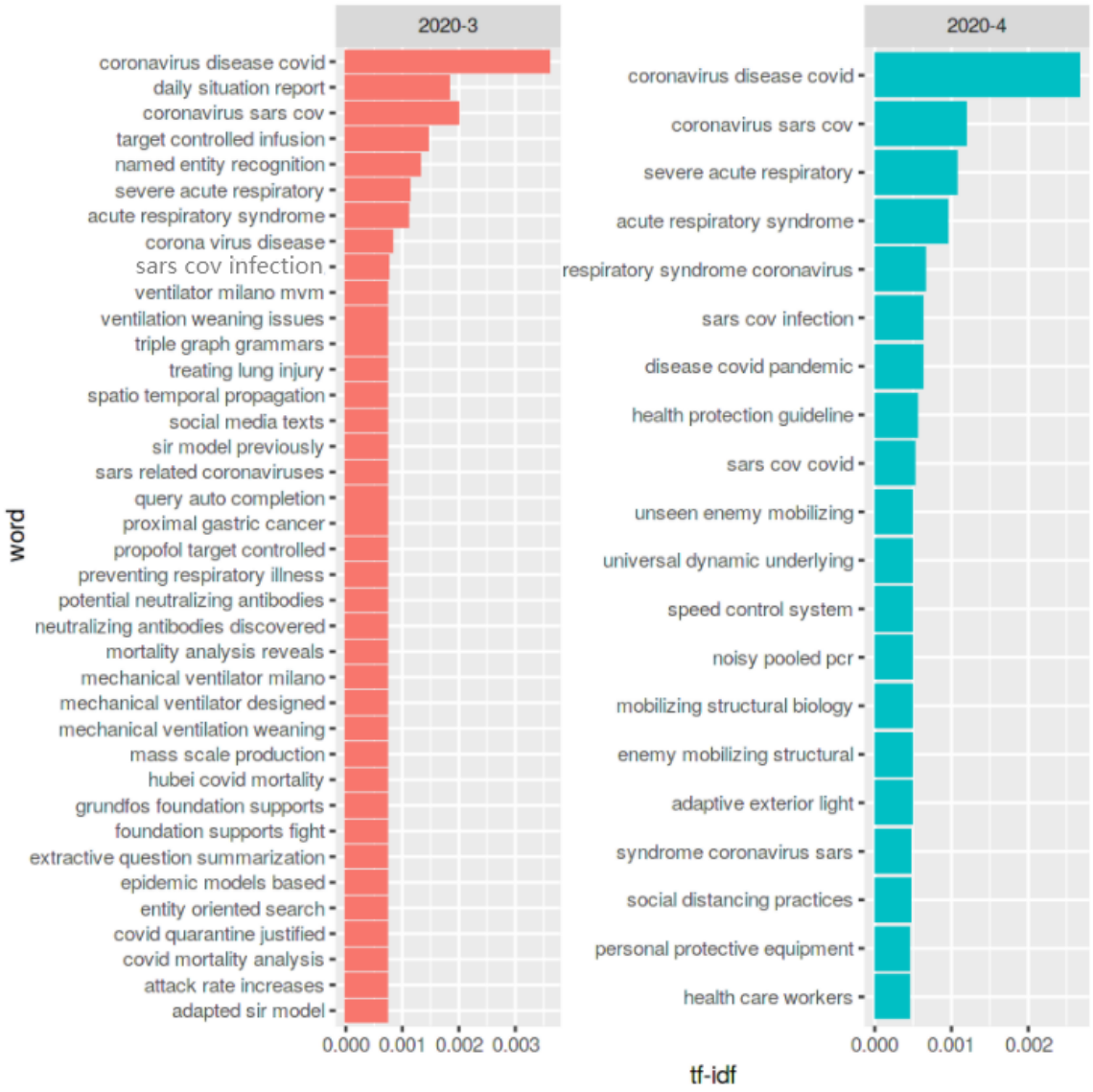

Figure 37: Trigram for March and April, 2020

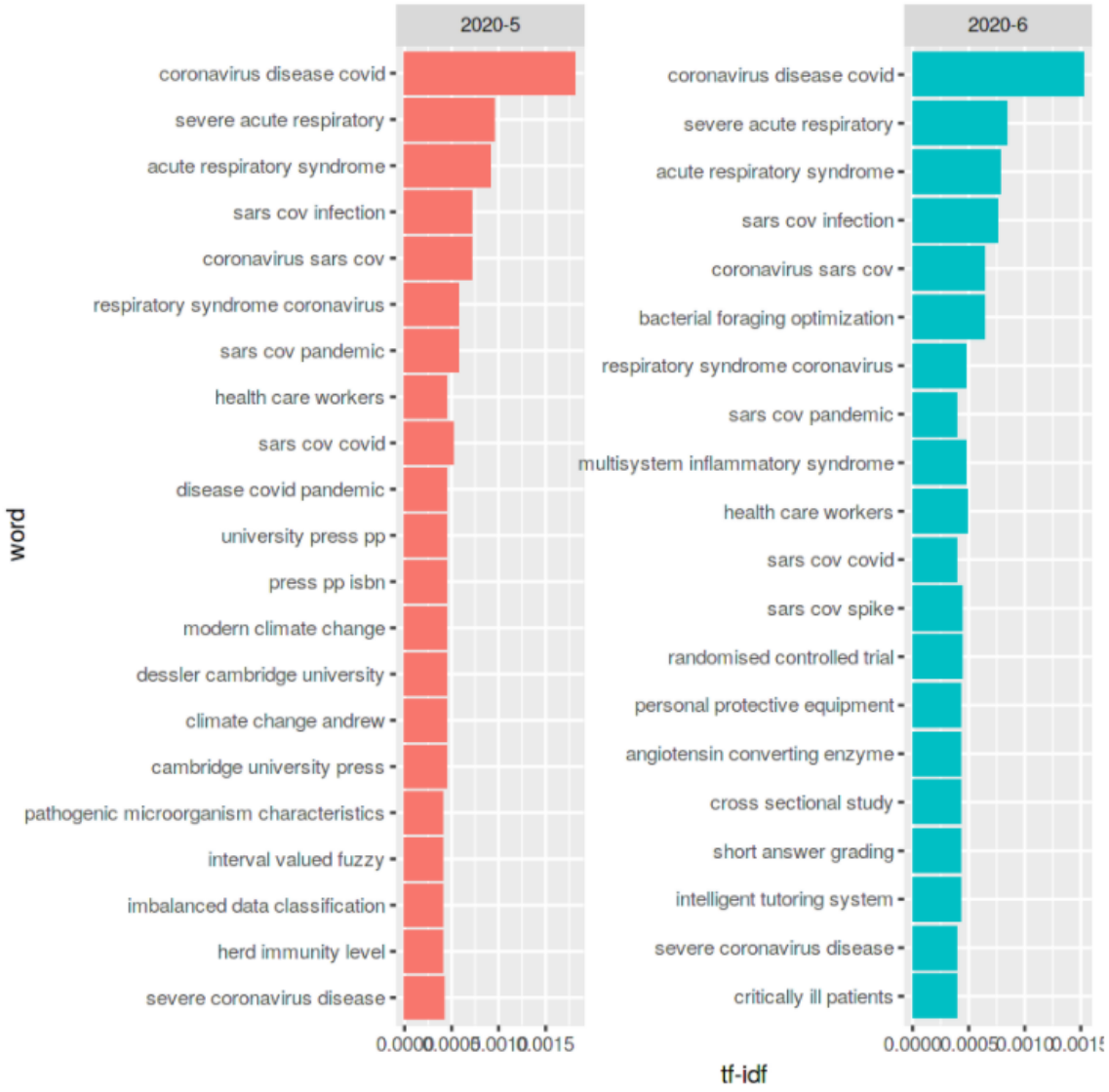

Figure 38: Trigram for May and June, 2020

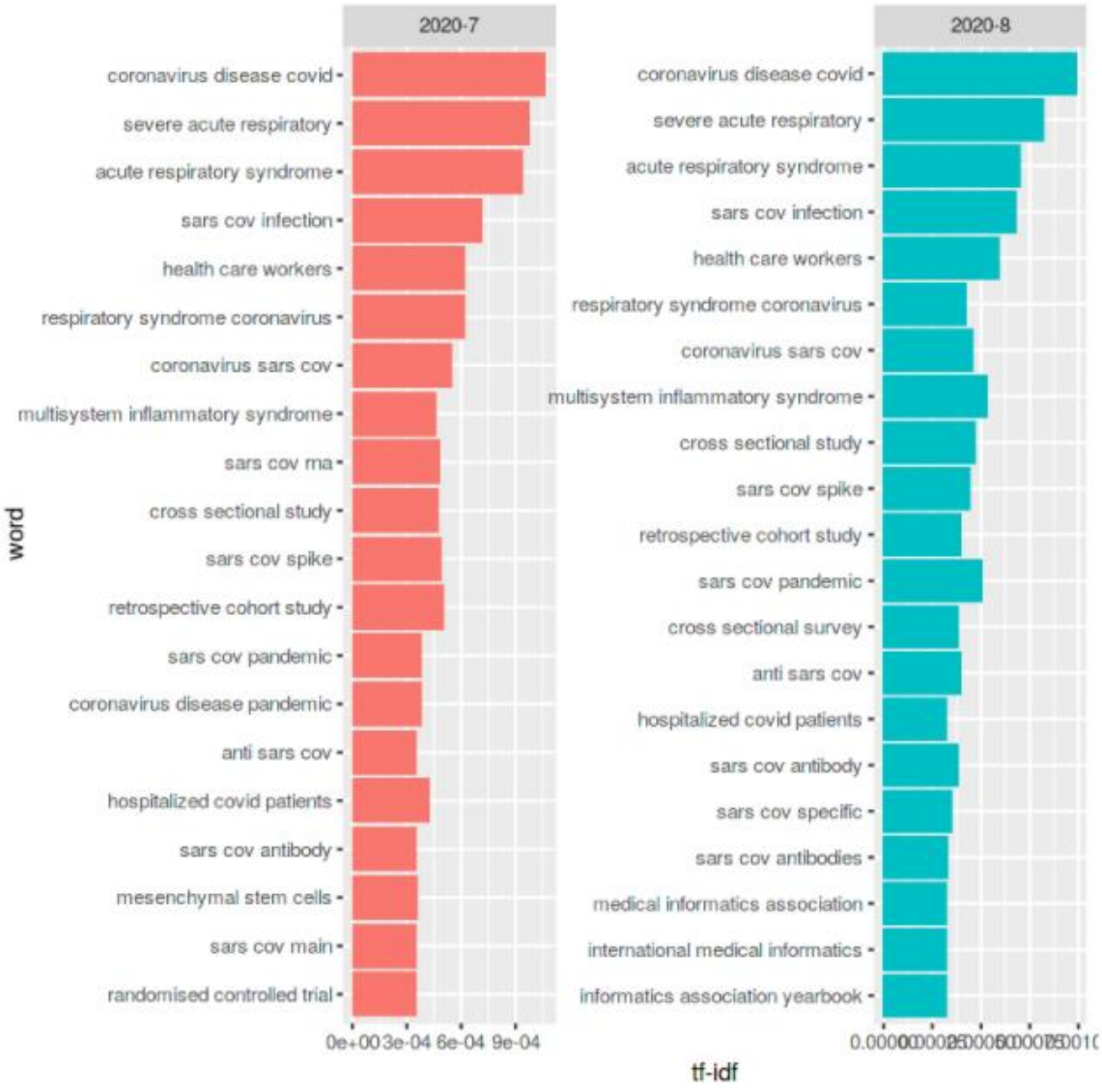

Figure 39: Trigram for July and August, 2020

In the same way as in Section 4.1, we calculate the frequency of individual words and report in Tables 9-11 the top ten words with their frequency and relative frequency, where the three high-frequency words are “Pandemic”, “Coronavirus” and “Health”.

Table 9 shows that "Pandemic COVID" appears 11132 times in the manuscripts, which is the most, and "Pandemic Health", "Pandemic Coronavirus" and "Care Pandemic" show at least 1000 times. Some key words (e.g., impact, disease, patients) combine with "Pandemic" at least 800 times.

Table 10 illustrates that “Disease Coronavirus” has been mentioned 5118 times in the CORD-19 dataset. The words for symptoms, for example, “Respiratory Coronavirus”, “Severe Coronavirus”, and “Syndrome Coronavirus”, display at least 1000 times.

Table 11 demonstrates that “Health COVID” appears 3300 times in the COVID-19 open research dataset, showing together with “Pandemic Health” in 1290 times. The words “Health Mental”, “Health Public”, and “Care Health” reveal key issues during the COVID-19 pandemic, which show at least 1000 times.

Table 9: The top 10 frequency words associated with "Pandemic"

Word 1	**Word 2**	**Times (n) (Proportion)**
Pandemic	COVID	11332 (10.01%)
Pandemic	Health	1290 (1.14%)
Pandemic	Coronavirus	1287 (1.14%)
Care	Pandemic	1240 (1.09%)
Impact	Pandemic	994 (0.88%)
Disease	Pandemic	949 (0.84%)
Patients	Pandemic	895 (0.79%)
Pandemic	Management	862 (0.76%)
Pandemic	SARS	738 (0.65%)
Pandemic	COV	726 (0.64%)

Table 10: The top 10 frequency words associated with "Coronavirus"

Word 1	**Word 2**	**Times (n) (Proportion)**
Disease	Coronavirus	5118 (5.65%)
COVID	Coronavirus	3785 (3.88%)
Patients	Coronavirus	1561 (1.60%)
Pandemic	Coronavirus	1287 (1.31%)
SARS	Coronavirus	1140 (1.17%)
Respiratory	Coronavirus	1103 (1.13%)
Severe	Coronavirus	1098 (1.24%)
Syndrome	Coronavirus	1044 (1.07%)
Acute	Coronavirus	1041 (1.07%)

Table 11: The top 10 frequency words associated with "Health"

Word 1	**Word 2**	**Times (n) (Proportion)**
Health	COVID	3300 (6.87%)
Pandemic	Health	1290 (2.69%)
Health	Mental	1113 (2.31%)
Health	Public	931 (1.93%)
Care	Health	923 (1.92%)
Health	Coronavirus	475 (0.99%)
health	Workers	441 (0.91%)
Study	Health	412 (0.86%)
Impact	Health	409 (0.85%)
Disease	Health	352 (0.73%)

5.2 N-gram Model

As in Chapter 4, we visualize how words appear together. We continue to show the structure of two words (i.e., Bigram) and three words (i.e., Trigram), and display them in Tables 12-13, respectively. It can be clearly seen from the results that the words related to "SARS COV" appear most frequently for 13,651 times. The number of occurrences of the phrases "COVID Pandemic" and "Coronavirus Disease" is similar to the results obtained from the TF-IDF model.

Table 12: The top 10 frequency words of the Bigram Model

Word 1	**Word 2**	**Times (n) (Proportion)**
SARS	COV	13651 (2.71%)
COVID	Pandemic	10141 (2.01%)
Coronavirus	Disease	5393 (1.07%)
Disease	COVID	2527 (0.50%)
COVID	Patients	2470 (0.49%)
Systematic	Review	2253 (0.45%)
COV	Infection	1873 (0.37%)
COVID	Outbreak	1762 (0.35%)
Meta	Analysis	1426 (0.28%)
Acute	Respiratory	1364 (0.27%)

Table 13: The top 10 frequency words of the Trigram Model

Word 1	**Word 2**	**Word 3**	**Times (n) (Proportion)**
Coronavirus	Disease	COVID	2368 (1.04%)
SARS	COV	Infection	1852 (0.81%)
Severe	Acute	Respiratory	916 (0.40%)
Respiratory	Syndrome	Coronavirus	844 (0.37%)
Acute	Respiratory	Syndrome	843 (0.37%)
Coronavirus	SARS	COV	672 (0.29%)
Cross	Sectional	Study	480 (0.21%)
SARS	COV	COVID	400 (0.18%)
SARS	COV	Spike	400 (0.18%)
Respiratory	Distress	Syndrome	337 (0.15%)

5.3 Topic Modeling and Visualization

5.3.1 Grid Search of the Optimal Number of Topics

Now we aim to find the most significant words in each cluster. The K-means method clusters the articles but does not label the topics. Through topic modeling we want to find out what the most important terms for each cluster are. This adds more meaning to the cluster by giving keywords to quickly identify the themes of the clusters.

In the same manner as in Section 4.4, we need to find the optimal number of topics first. In Figure 40, the highest coherence score is 0.5489, yielding about 17 topics.

Then, as shown in Figure 41, when $K = 17$, *PyLDAvis* bubbles are fairly evenly distributed in the quadrant with less overlap, which shows that 17 is the optimal number of topics we are looking for.

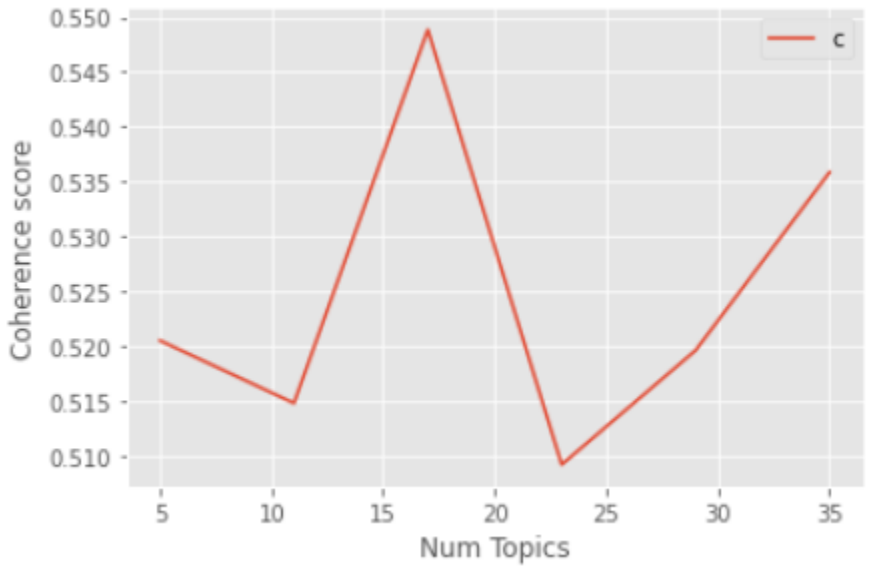

Figure 40: Coherence Score for Optimal Number of Topics

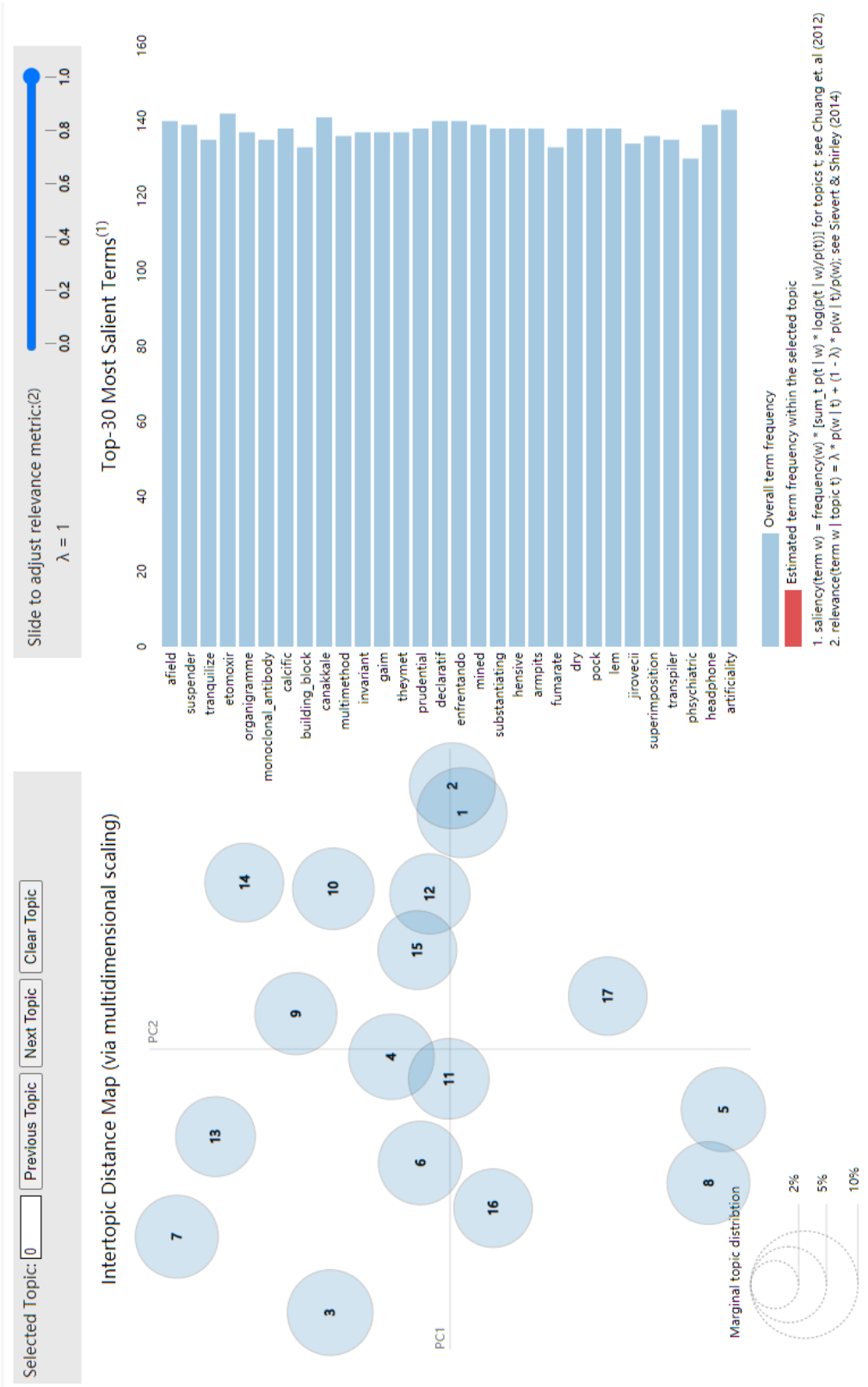

Figure 41: pyLDAvis with 17 topics

5.3.2 Visualization with Word Cloud

We conducted a word cloud analysis of each topic. A maximum word limit of 10 and 50 was set, respectively. Common English words were removed. Below is the list of those identified 17 topics:

- Topic #0: Vaccine

Topic #0 talks about vaccine. There are some dominated words in Figure 42, such as “immune”, “vaccine”, “pathology”. Those words can be summarized as related to vaccine.

Figure 42: Word Cloud Topic #0

- Topic #1: Hospital

Figure 43 shows the word cloud of topic #1. We find that the word “pandemic”, “hospital” and “patient” are the most dominant. Other words like “emergency”, “surgery”, “medical” and “healthcare” indicate the relevance to hospital.

Figure 43: Word Cloud Topic #1

- Topic #2: COVID-19 research study

The word cloud visualization of topic #2 is shown by Figure 44. The words “study”, “database”, “analysis” and “result” are dominating in this topic segment. In addition, words related to research study include “analysis”, “evidence”, “database”, “method”.

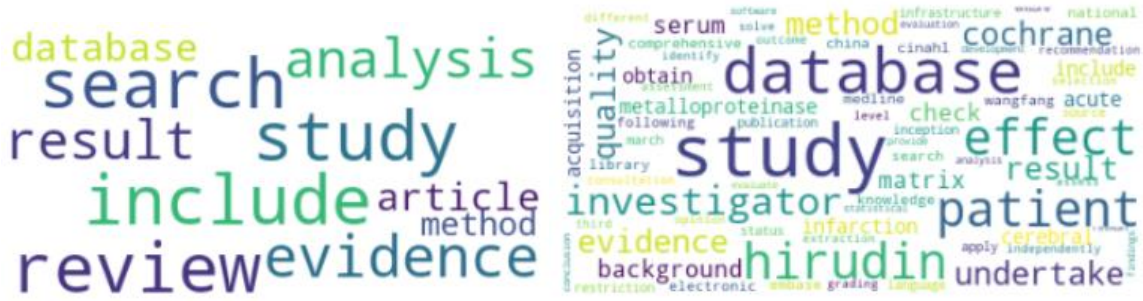

Figure 44: Word Cloud Topic #2

- Topic #3: Treatment

Topic #3 is about the treatment of COVID-19, which can be seen from the terms “therapy”, “clinical”, “control”, “drug”, and “trial”. In addition, the word “hydroxychloroquine” points out that researchers are demonstrating whether this medicine is useful. The word cloud of topic #3 can be seen in Figure 45.

Figure 45: Word Cloud Topic #3

- Topic #4: Effects on the brain and nerves

Figure 46 shows the word cloud of topic #4. From the terms in the word cloud, it can be concluded that topic #4 is about the effects on the brain and nerves, which can be seen

from the terms “nerve”, “brain”, “disorder” and “cognitive”. Topic #4 talks about the COVID-19 impacts on people's brains. Also, the term "smoking" may indicate that whether the COVID-19 patient smokes or not.

Figure 46: Word Cloud Topic #4

- Topic #5: Study comparison

According to the word cloud in Figure 47, topic #5 is about the studies that compare the COVID-19 results and outcomes, suggesting by the words “compare”, “result” and “outcome” as well as “baseline”, “initial”, and “severity”.

Figure 47: Word Cloud Topic #5

- Topic #6: Pregnant woman

The words in topic #6 concern the impact of COVID-19 on pregnant women, which can be seen from three dominant words in the word cloud, “pregnant”, “woman” and “pediatric”. The word cloud of topic #6 is shown in Figure 48.

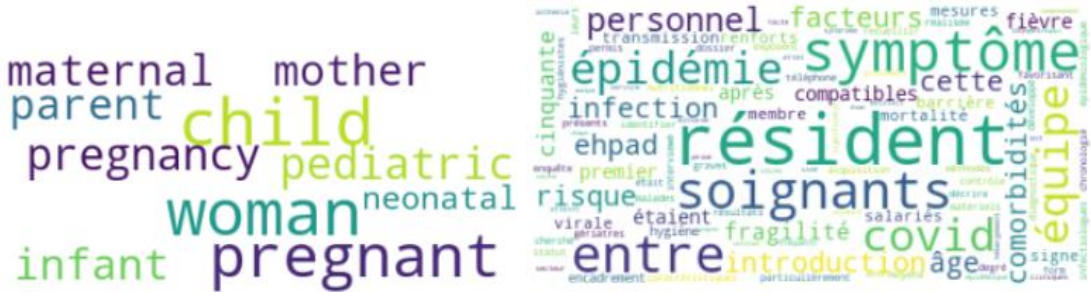

Figure 48: Word Cloud Topic #6

- Topic #7: Respiratory

Topic #7 is related to the number of cases for COVID-19 respiratory, from Figure 49, which can be seen from three dominant words in its word cloud, “infection”, “respiratory”, and “acute”.

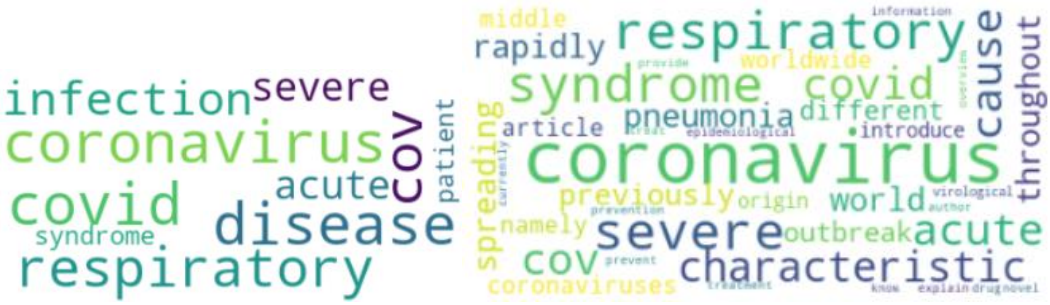

Figure 49: Word Cloud Topic #7

- Topic #8: Application

Topic #8, in Figure 50, talks about the application related to COVID-19. The words “network”, “prediction”, “analysis” and “information” may indicate the researchers use the Neural Networks approach to analyze information.

Figure 50: Word Cloud Topic #8

- Topic #9: Education

According to the word cloud in Figure 51, topic #9 is about student life during the COVID-19 pandemic. The words “learning”, “online”, “education” and “school” indicate the topic about education.

Figure 51: Word Cloud Topic #9

- Topic #10: Masks

Topic #10, shown in Figure 52, is related to masks. Researchers use aerosols to test whether the virus can be transmitted through masks. The words “ultrafine”, “fabric”, “mask”, “protective” and “transmission” indicate the topic about masks.

Figure 52: Word Cloud Topic #10

- Topic #11: Public Response

The words in topic #11 suggest the public response to the COVID-19 pandemic, which can be seen from three dominant words in the word cloud, “crisis”, “ministry” and “secretariat”. The word cloud of topic #11 is shown in Figure 53.

Figure 53: Word Cloud Topic #11

- Topic #12: Outbreak in China

Topic #12 is about the COVID-19 outbreak in China. It can be seen from the three dominant words in the word cloud, “China”, “outbreak” and “infection”. In addition, the words “incubation”, “asymptomatic” and “contagiousness” might be related to researchers who use Chinese COVID-19 data to study the virus. The word cloud of topic #12 is shown by Figure 54.

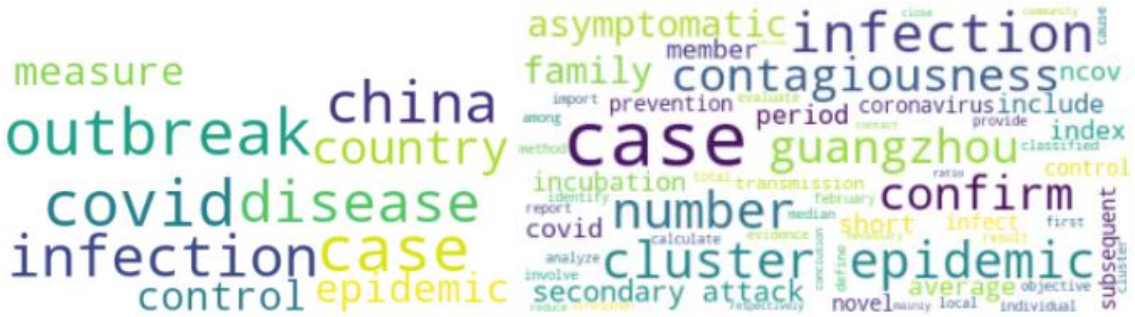

Figure 54: Word Cloud Topic #12

- Topic #13: Social distancing

Topic #13 might talk about the study on social distancing. From Figure 55, the words "social", "distance", "reduce" and "transmission" may indicate the topic about researchers who study whether or not social distancing is effective.

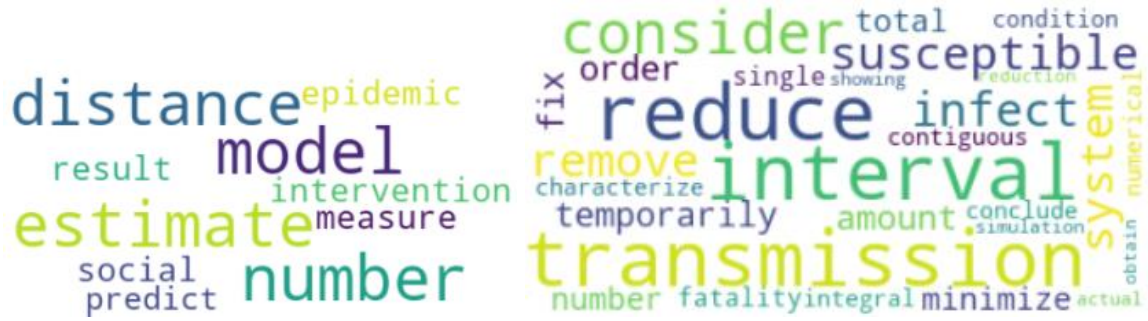

Figure 55: Word Cloud Topic #13

- Topic #14: Psychology

The topic discussed in topic #14 is about psychology, which can be seen from the terms "psychological", "mental", "anxiety". In addition, the words "survey", "gender", "mental" may suggest that researchers used survey to investigate and compare mental health of different gender. The word cloud of topic #14 can be seen in Figure 56.

Figure 56: Word Cloud Topic #14

- Topic #15: Economic, Pollution and Population

Topic #15 shows the impact of COVID-19 on economy, pollution and population in Figure 57. The terms “increase”, “mortality”, “population” and “reduction” might show the impact of COVID-19. Also, the words “emission” and “pollution” may show the connection of the lockdown and pollution emissions.

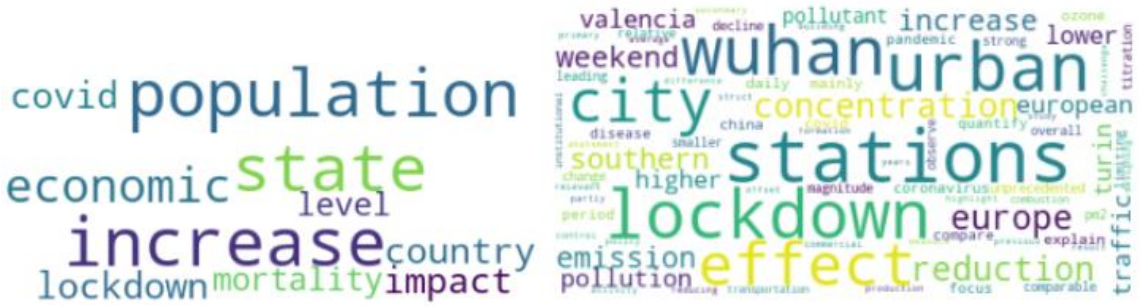

Figure 57: Word Cloud Topic #15

- Topic #16: Biology

Topic #16, in Figure 58, talks about using biology methods to study COVID-19. Some proprietary words “protein”, “molecule”, “cleavage”, “enzyme” may indicate how researchers explore the composition of viruses.

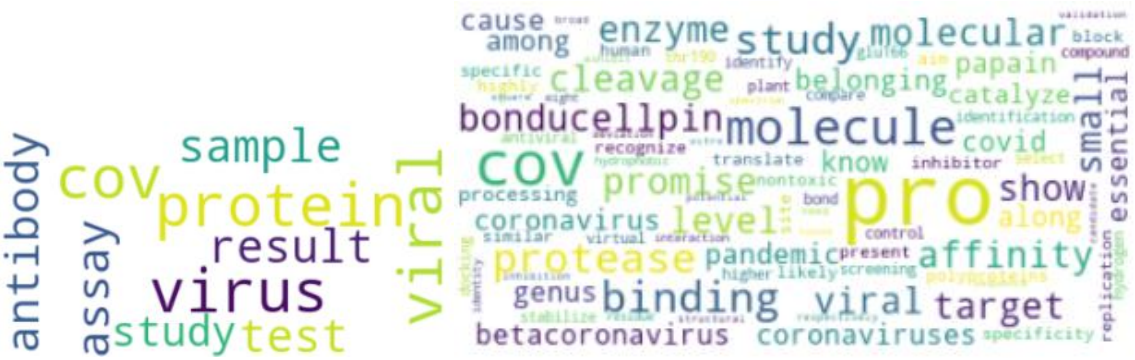

Figure 58: Word Cloud Topic #16

5.4 Findings and Discussions

In this chapter, we obtain the analysis results from COVID-19 Open Research Dataset by using three standard models: TF-IDF model, LDA model, and our proposed model. From the results we may gain understanding about the focus of the published articles for the period of January 2020 to August 2020. From these text results, it seemed that the focus of researchers and scholars is more on the treatment, symptoms and vaccines of COVID-19. For example, how COVID-19 may affect people's immune system and various organs, what the effective treatment is, and how to effectively control the spread of the virus based on data and analysis.

Chapter 6 Evaluation

In this Chapter, we evaluate and compare the model performance. In Section 6.1, we show the numerical results which involve 4 models: LDA model, Topical N-Gram model, 2NN Topical N-gram LDA model, and 3NN Topical N-gram. In Section 6.2, we do a qualitative analysis which shows the result of topic words with different models.

6.1 Evaluation Results

Perplexity is an intrinsic evaluation metric which is widely used for the language model evaluation. It is particularly useful to evaluate the topic model performance. To measure the modeling power, perplexity calculates the inverse loglikelihood of the unobserved documents. The lower perplexity value means the better models with less uncertainties about the unobserved documents. The perplexity is presented as:

$$perplexity(D_{text}) = exp\left\{-\frac{\sum_{d=1} log\, p(w_d)}{\sum_{d=1} N_d}\right\},$$

where w_d represents the number of words in the document d, N_d is the length of the document d, and $log\, p(w_d)$ is calculated as log-likelihood of each unobserved document.

Table 14 shows the perplexity results for the models. For the small dataset, CBC news dataset, the 2NN TNG performs better than other models, giving the lowest score -8.589. But for the large dataset, CORD-19, LDA model yields the lowest score -8.497, but the result of 2NN TNG is very close to that of the LDA model. These results seem to suggest that for small datasets, our model performs better and has less perplexity, but performs not as well as LDA for big datasets.

Table 14: Perplexity Results of Various Model

Dataset	LDA	Topical N-gram	2NN Topical N-gram Model	3NN Topical N-gram Model
CBC News	-7.5218	-8.0284	**-8.589**	-8.2065
CORD-19	**-8.497**	-8.0702	-8.5626	-8.2104

The concept of topic coherence combines a number of measures into a framework to evaluate the coherence between topics inferred by a model. It can be used for determining the optimal number of topics. Table 15 shows that the coherence score of 2NN TNG is higher than that of other models in both datasets. This may indicate that our model performs well in both large and small corpus topic modeling and can provide a more accurate optimal number of topics.

Table 15: Coherence Score of Various Model

Dataset	LDA	Topical N-gram	2NN Topical N-gram Model	3NN Topical N-gram Model
CBC News	0.321	0.222	**0.325**	0.282
CORD-19	0.548	0.532	**0.568**	0.452

Table 16 shows the computation time required by each model. The table shows that 2NN TNG is better than 3NN TNG and LDA in terms of the time required to calculate the topic, our proposed model spends the same time in a small corpus as LDA, but it takes less time for a large corpus.

Table 16: Working Time of Models (in Minutes) (RAM 12GB)

Dataset	LDA	Topical N-gram	2NN Topical N-gram Model	3NN Topical N-gram Model
CBC News	**18**	20	**18**	30
CORD-19	22	34	**21**	63

6.2 Qualitative Analysis

We choose the dataset COVID-19 open research (CORD-19) and show the top topic words of Topical N-gram model (TNG) and Deep Topical N-gram model (DTNG) in Table 17 and Table 18. At first sight, there are many common words in both tables (e.g., “COVID-19”, “outbreak”, “diseases”), which are closely related to the category of dataset “CORD-19” and the words we show in the below part.

Table 17: Top topic words discovered by TNG

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
high risk	severe acute	social distancing	covid19 case	social medium
patient method	patient covid19	mental health	chain reaction	syndrome coronavirus
care system	coronavirus sarscov2	public health	care unit	symptom onset
immune system	respiratory failure	**result show**	disease severity	cytokine storm
critically ill	aim study	important role	new coronavirus	higher risk

Table 18: Top topic words discovered by DTNG

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
immune system	clinical feature	social distancing	case series	world health
high risk	hospitalized patient	public health	chain reaction	clinical characteristic
fatality rate	distress syndrome	mental health	fever cough	clinical trial
personal protective	severe acute	**health organization**	respiratory infection	social medium
tested positive	respiratory failure	outbreak covid19	patient underwent	mean age

However, when we carefully compare the two tables, there exists some differences. In Table 17, there are some normal words which have little help for topic clustering (e.g., “patient method”, “care system”, “coronavirus sarscov2”), but they still get large weights and appear in the front of topic words list. Obviously, DTNG misinterprets these words because they usually appear together with the real topic words. In Table 18, we are not able to find them in top words list, because extra information from Deep neural network makes a timely supplement. Besides, similar situations also occur for words “immune system” and “result show”. The word “immune system” is an apparent topic word, so DTNG gives it a larger weight in topic words list. For the word “result show”, as a no means two words, it can provide less topic information than a phrase, so DTNG decreases its weight and put “health organization” in the second place. From the above analysis we can see that DTNG tends to show a better topic clustering ability than TNG.

Chapter 7 Summary

7.1 Conclusion

In this thesis, we have presented the Topic modelling technique, Latent Dirichlet Allocation with a N-gram model and Deep neural network variant Deep Topical N-gram model. The proposed model shows good performance, suggested by our data analyses. Topics were learned from the datasets COVID-19 News Articles Open Research Dataset and COVID-19 Open Research Dataset. Tables 14-18 indicate that the proposed model is superior to the existing topic model in terms of reducing computation time and the ability of handing large data.

We hope that the analyses in this thesis shed light on learning the COVID-19 pandemic and help gain a better understanding of the concerns and needs of people with respect to COVID-19 related issues.

7.2 Future Work

A possible future exploration may be carried out by applying our methods to the other text classification tasks, such as PubMed data, in particular, the usage of fastText [60], which provides pre-trained word embedding generated with word2vec [61] using a considerable amount of texts. Additionally, the exploration of other text representation extensions of word2vec can also be a future research project. Also, our proposed model can provide a possible basis for developing deep neural networking based variant of Biterm Topic Model (BTM), which is a topic modeling technique for short texts messages.

References or Bibliography

1. Wu, Y. C., Chen, C. S., & Chan, Y. J. (2020). The outbreak of COVID-19: an overview. *Journal of the Chinese medical association*, 83(3), 217.

2. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y. & Yu, T. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The lancet*, 395(10223), 507-513.

3. Welcome - COVID-19-Canada. (2020). Retrieved July 20, 2020, from https://covid-19-canada.uwo.ca/

4. Velavan, T. P., & Meyer, C. G. (2020). The COVID‐19 epidemic. Tropical medicine & international health, 25(3), 278.

5. Canada, P. H. A. (2020). Coronavirus disease (COVID-19): Outbreak update. *Canada.ca*. https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html.

6. Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, K., Tutubalina, E. & Chowell, G. (2020). A large-scale COVID-19 Twitter chatter dataset for open scientific research-an international collaboration. *arXiv preprint arXiv:2004.03688*.

7. Ruz, G. A., Henrquez, P. A., & Mascareo, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. *Future Generation Computer Systems*, 106, 92-104.

8. Kousha, K., & Thelwall, M. (2020). COVID-19 publications: Database coverage, citations, readers, tweets, news, Facebook walls, Reddit posts. *Quantitative Science Studies*, *1*(3), 1068-1091.

9. Müller, M., Salathé, M., & Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

10. Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *Plos one*, *16*(2), e0245909.

11. Zhang, J. S., Keegan, B. C., Lv, Q., & Tan, C. (2020). A tale of two communities: Characterizing reddit response to covid-19 through/r/china flu and/r/coronavirus. *arXiv preprint arXiv:2006.04816*.

12. Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E.W. & Baddour, K. (2020). Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, 12(3).

13. Kreps, S. E., & Kriner, D. (2020). Medical misinformation in the COVID-19 pandemic. *Available at SSRN 3624510*.

14. Zheng, N., Du, S., Wang, J., Zhang, H., Cui, W., Kang, Z., Yang, T., Lou, B., Chi, Y., Long, H. & Xin, J. (2020). Predicting COVID-19 in China using hybrid AI model. *IEEE transactions on cybernetics*, 50(7), 2891-2904.

15. Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10), 2733-2742.

16. Hosseini, P., Hosseini, P., & Broniatowski, D. A. (2020). Content analysis of Persian/Farsi Tweets during COVID-19 pandemic in Iran using NLP. *arXiv preprint arXiv:2005.08400*.

17. Kolluri, J., Razia, S., & Nayak, S. R. (2020). Text classification using Machine Learning and Deep Learning Models. *Available at SSRN 3618895*.

18. Singh, B., Yadav, I., Agarwal, S., & Prasad, R. (2009). An efficient word searching algorithm through splitting and hashing the offline text. In *2009 International Conference on Advances in Recent Technologies in Communication and Computing, 387-389*. IEEE.

19. Wielinga, B. J., Schreiber, A. T., & Breuker, J. A. (1992). KADS: A modelling approach to knowledge engineering. *Knowledge acquisition*, 4(1), 5-53.

20. Zhou, L., Wang, L., Ge, X., & Shi, Q. (2010). A clustering-Based KNN improved algorithm CLKNN for text classification. In *2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2010), Vol. 3, 212-215*. IEEE.

21. Diday, E., Govaert, G., Lechevallier, Y., & Sidi, J. (1981). Clustering in pattern recognition. *In Digital Image Processing, 19-58.* Springer, Dordrecht.

22. Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based framework for text categorization. *Procedia Engineering*, 69, 1356-1364.

23. Uther, W., Mladenić, D., Ciaramita, M., Berendt, B., Kołcz, A., & Grobelnik, M. et al. (2011). TF–IDF. *Encyclopedia of Machine Learning*, 986-987.

24. Miltsakaki, E., & Kukich, K. (2004). Evaluation of text coherence for electronic essay scoring systems. Natural Language Engineering, 10(1), 25.

25. Joshi, K. D., & Nalwade, P. S. (2013). Modified k-means for better initial cluster centres. *International Journal of Computer Science and Mobile Computing*, 2(7), 219-223.

26. Ding, C., Li, T., & Peng, W. (2006, August). NMF and PLSI: Equivalence and a Hybrid Algorithm. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 641-642.

27. Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*, 1445-1456.

28. Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 161-169.

29. Han, R. (2020). COVID-19 News Articles Open Research Dataset. Kaggle. https://www.kaggle.com/ryanxjhan/cbc-news-coronavirus-articles-march-26.

30. CBC/Radio Canada. (2020). CBCnews. https://www.cbc.ca/search?q=coronavirus+news§ion=all&sortOrder=relevance&media=all.

31. SpaCy. Industrial-strength Natural Language Processing in Python. (2020). Retrieved from https://spacy.io/

32. McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14(9).

33. Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1), 188-230.

34. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825-2830.

35. Singh, H., & Kaur, K. (2013). New method for finding initial cluster centroids in K-means algorithm. *International Journal of Computer Applications*, 74(6).

36. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

37. Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in bioinformatics*, 15(5), 788-797.

38. Vargiu, E., & Urru, M. (2013). Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Artif. Intell*. Research, 2(1), 44-54.

39. Prakash, M., & Rashid, D. (2017). A review of programming languages for web scraping from software repository sites. *International Journal of Engineering and Technology*, 9(3), 2383-2388.

40. Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W. & Mooney, P. (2020). *Cord-19: The covid-19 open research dataset*. ArXiv.

41. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3, 993-1022.

42. Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86.

43. Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

44. Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

45. Bellegarda, J. R. (1997). A latent semantic analysis framework for large-span language modeling. *In Fifth European Conference on Speech Communication and Technology*.

46. Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37-52.

47. Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE International Conference on Data Mining* (ICDM 2007) , 697-702, IEEE.

48. Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4), 43-52.

49. Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1), S63-S63.

50. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

51. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of The North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480-1489.

52. Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

53. Ma, Y., Peng, H., & Cambria, E. (2018). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

55. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

56. Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, 977-984.

57. Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2), 211.

58. Hu, W., Shimizu, N., Nakagawa, H., & Sheng, H. (2008). Modeling Chinese documents with topical word-character models. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 345-352.

59. Johnson, M. (2010). PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1148-1157.

60. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

61. Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, 23(1), 155-162.

62. Ghosh, S. (2018). Topic modelling with Latent Dirichlet Allocation (LDA) in Pyspark. *Medium*. https://medium.com/@connectwithghosh/topic-modelling-with-latent-dirichlet-allocation-lda-in-pyspark-2cb3ebd5678e.

63. SysNucleus. (2020). WebHarvy Web Scraper. *Web Scraping Explained*. https://www.webharvy.com/articles/what-is-web-scraping.html.

Curriculum Vitae

Name: Yuan Du

Post-secondary Education and Degrees:
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
2011-2015 BS(CS).

Honours and Awards:
Western Graduate Research Scholarship
2019-2020

Related Work Experience
Teaching Assistant
University of Western Ontario
2019-2020

Publication:

Liu, D., Du, Y., Charvadeh, Y.K., Cui, J., Chen, L.P., Deng, G., Zhang, Q., Cai, K., He, J., He, W. and Yi, G.Y., (2020). A real time and interactive web-based platform for visualizing and analyzing COVID-19 in Canada. *International Journal of Statistics and Probability*, 9(5), 23-29.